

# **Identification of Novel Causative Genes for Colorectal Adenomatous Polyposis**

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

**Sukanya Horpaopan**

aus

**Nakhonsawan, Thailand**

Bonn, 2015

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät  
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Die vorliegende Arbeit wurde am Institut für Humangenetik  
der Rheinischen Friedrich-Wilhelms-Universität zu Bonn angefertigt.

1. Gutachter: Prof. Dr. Stefan Aretz
2. Gutachter: Prof. Dr. Michael Hoch

Tag der Promotion: February 9, 2015  
Erscheinungsjahr: 2015

# TABLE OF CONTENTS

	Page
Acknowledgements .....	VI
Declaration .....	VII
List of abbreviations .....	VIII
1. INTRODUCTION .....	1
2. BASIC PRINCIPLES .....	3
2.1. Hereditary Colorectal Cancer .....	3
2.1.1. Differential diagnosis .....	4
2.1.2. Adenoma-carcinoma sequence .....	5
2.1.3. Knudson two-hit hypothesis .....	6
2.2. Genetics of adenomatous polyposis syndromes .....	7
2.2.1. Familial adenomatous polyposis (FAP) .....	7
2.2.2. <i>MUTYH</i> -associated polyposis (MAP) .....	9
2.2.3. Mutation negative adenomatous polyposis .....	10
2.3. Human genome variation .....	11
2.3.1. Single nucleotide polymorphisms (SNPs) .....	12
2.3.2. Tandem repeats .....	12
2.3.3. Structural variations .....	12
2.4. Identification of causative genes in human disease .....	18
2.4.1. Homozygosity mapping .....	19
2.4.2. Loss of heterozygosity (LOH) analysis .....	20
2.4.3. Linkage analysis .....	21
2.4.4. Genome-wide association study (GWAS) .....	22
2.4.5. Copy number variation (CNV) analysis .....	23
2.4.6. High-throughput sequencing .....	26
2.5. Validation of functionally relevant candidate genes .....	27
2.5.1. Recurrent findings .....	27
2.5.2. Segregation analysis .....	27
2.5.3. Gene expression in relevant target tissues .....	28
2.5.4. Candidate gene approach .....	28
2.5.5. Pathway enrichment analysis/network analysis .....	28
2.6. Scope of the thesis .....	30
3. MATERIALS AND METHODS .....	31

3.1. Databases.....	31
3.2. Devices.....	32
3.3. Software .....	33
3.4. Commercial reagents.....	34
3.5. Study samples .....	35
3.5.1. Initial patient cohort .....	35
3.5.2. NGS validation cohort.....	36
3.5.3. Heinz Nixdorf RECALL (HNR) study controls .....	36
3.5.4. GWAS replication study.....	37
3.6. DNA and RNA preparations .....	37
3.6.1. DNA extraction using desalting method .....	37
3.6.2. Formalin fixed paraffin embedded (FFPE) tissue DNA isolation.....	38
3.6.3. RNA extraction using the PAX gene kit.....	38
3.6.4. Determination of concentration and quality .....	39
3.6.5. First-strand cDNA synthesis .....	39
3.7. Polymerase Chain Reaction (PCR) .....	40
3.7.1. Basic principle .....	40
3.7.2. Primer design .....	40
3.7.3. PCR reaction components.....	41
3.7.4. Cycling step.....	42
3.7.5. Agarose gel electrophoresis .....	42
3.7.6. PCR product purification.....	42
3.8. Sanger sequencing .....	43
3.8.1. Basic principle .....	43
3.8.2. Reaction components.....	44
3.8.3. Cycling step.....	44
3.8.4. Cycle sequencing product cleaning .....	44
3.8.5. Capillary electrophoresis .....	45
3.9. APC transcript analysis .....	45
3.9.1. Primer design .....	45
3.9.2. cDNA analysis .....	46
3.9.3. Sanger sequencing.....	46
3.9.4. Data analysis .....	46
3.9.5. Genomic DNA analysis.....	47
3.9.6. Haplotype analysis .....	47
3.9.7. In-silico analysis .....	47

3.10. Genome-wide SNP array hybridization .....	47
3.10.1. Genotyping based on BeadArray Technology (Illumina®).....	47
3.10.2. Protocol .....	48
3.10.3. Bead decoding .....	49
3.10.4. Quality control of raw data .....	49
3.11. Identification of putative CNVs .....	50
3.11.1. Final reports .....	50
3.11.2. CNV calling.....	50
3.12. CNV analysis .....	52
3.12.1. Known candidate gene survey.....	52
3.12.2. Filtering CNVs .....	52
3.13. CNV validation .....	57
3.13.1. Quantitative PCR (qPCR) .....	57
3.13.2. CNV validation by qPCR using SYBR Green I.....	57
3.13.3. Data analysis.....	59
3.13.4. Copy number calculation ( $2^{-\Delta\Delta CT}$ method).....	59
3.14. Co-segregation analysis.....	60
3.15. Gene expression analysis .....	60
3.15.1. Gene expression in human colon cDNA .....	60
3.15.2. PCR and agarose gel electrophoresis .....	61
3.16. Network analysis.....	62
3.17. Candidate gene prioritization.....	62
3.17.1. Frequency of finding .....	62
3.17.2. Segregation analysis .....	62
3.17.3. Data mining .....	63
3.18. TaqMan® gene expression analysis .....	63
3.18.1. Basic principle .....	63
3.18.2. Relative quantitative PCR (RT-PCR) .....	64
3.18.3. Reaction components.....	64
3.18.4. Cycling step.....	65
3.18.5. Data analysis.....	65
3.19. Targeted next generation sequencing .....	66
3.19.1. Basic principle .....	66
3.19.2. Library preparation, target enrichment, and sequencing .....	67
3.19.3. Alignment, genotype calling, and variant annotation .....	67
3.19.4. Data analysis and filter .....	67

3.19.5. Validation of results .....	68
3.20. Genotyping based on MassExtend Reaction (Sequenom®).....	68
3.20.1. Basic principle .....	68
3.20.2. Selection of the genotyped SNPs .....	68
3.20.3. DNA preparation.....	70
3.20.4. Assay and primer design .....	70
3.20.5. PCR step .....	70
3.20.6. Digestion step.....	71
3.20.7. Extension primer adjustment .....	71
3.20.8. Extension step .....	71
3.20.9. Clean up reaction .....	72
3.20.10. Dispensing DNA on a chip.....	73
3.20.11. Mass spectrometry .....	73
3.20.12. Data analysis.....	73
3.21. TaqMan® SNP genotyping/allelic discrimination .....	73
3.21.1. Basic principle .....	73
3.21.2. DNA preparation.....	74
3.21.3. Primer and probe design .....	74
3.21.4. PCR step .....	74
3.21.5. Allelic discrimination data analysis.....	75
4. RESULTS.....	76
4.1. Transcript analysis of the <i>APC</i> gene .....	76
4.1.1. Agarose gel electrophoresis .....	76
4.1.2. Sanger sequencing of aberrant transcripts and genomic DNAs .....	77
4.1.3. In-silico analysis .....	80
4.1.4. Haplotype analysis .....	80
4.2. CNV analysis .....	82
4.2.1. Quality control of SNP array hybridization .....	82
4.2.2. CNV calling.....	82
4.2.3. Survey of CNV in known candidate genes.....	84
4.2.4. CNV filtering .....	86
4.2.5. CNV validation by qPCR using SYBR Green I.....	88
4.2.6. Co-segregation analysis .....	88
4.3. Candidate gene prioritization.....	91
4.3.1. Genes covered by the validated CNVs .....	91
4.3.2. Gene expression in human colon cDNA .....	92

4.3.3. Network and pathway analysis .....	93
4.3.4. Data mining .....	100
4.4. TaqMan® gene expression analysis of <i>CTNNB1</i> and <i>MUTYH</i> .....	104
4.5. Resequencing candidate genes .....	105
4.5.1. Sanger sequencing of <i>LZTFL1</i> .....	105
4.5.2. High throughput resequencing of remaining candidate genes .....	106
4.6. Screening for somatic point mutation .....	114
4.7. Replication of the GWAS performed in adenomatous polyposis.....	115
5. DISCUSSION.....	116
5.1. Transcript analysis of the <i>APC</i> gene .....	116
5.2. Novel causative gene identification .....	117
5.2.1. CNV analysis.....	117
5.2.2. Candidate gene prioritization .....	122
5.2.3. Validation of the clinical relevance of the candidate genes .....	127
5.3. Limitations of the study .....	130
6. SUMMARY .....	131
7. OUTLOOK/PERSPECTIVE .....	133
8. REFERENCES.....	135
List of publications.....	152
Appendices .....	153

# Acknowledgements

Firstly, I would like to express my gratitude to the German Academic Exchange Service (DAAD), which provided the financial support for my study in Germany and to Naresuan University, Thailand, for approving my study leave.

I would like to express my sincere appreciation to Dr. Alexander Zink, Dr. Per Hoffmann, Dr. Isabel Spier, all lab technicians particularly, Siegfried Uhlhaas, and to all members of the Institute of Human Genetics, University of Bonn, both at the Biomedical Center (BMZ) and the Department of Genomics, Life & Brain Center, who supported me during my studies. I thank Prof. Dr. Markus Nöthen who gave me the opportunity to perform my thesis at this place. I am realized how very lucky I am to be here and to do my PhD here.

I would like to extend my gratitude to Prof. Jurg Ott, Prof. Gavin Reynolds, to all collaborators; Cologne Center for Genomics (CCG), Center of Medical Genetics (MGZ) Munich, Prof. Dr. Holger Fröhlich, Dr. Holger Thiele, and to all patients who participate in the studies.

I am grateful to all committee members who have provided guidance and feedback on my works: Prof. Dr. Stefan Aretz, Prof. Dr. Michael Hoch, Prof. Dr. Walter Witke, and Prof. Dr. Sven Perner

I am thankful for friendship from everyone. My life here is joyful because of all of you.

I would like to thank my parents and my beloved sisters for your unconditional love, unlimited support, understanding, and for always being beside me.

Finally, I wish to express my deepest and sincere thanks to my supervisor, Prof. Dr. Stefan Aretz, for all kind advice, encouragement, being supportive in working and living, and a nice inspiration.



# Declaration

I, hereby confirm that this work is my own. This thesis has been written independently and with no other sources and aids than stated. The presented thesis has not been submitted to another university and I have not applied for a doctorate procedure so far.

Hiermit versichere ich, dass die vorgelegte Arbeit – abgesehen von den ausdrücklich bezeichneten Hilfsmitteln – persönlich, selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde. Aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle kenntlich gemacht worden.

Die vorliegende Arbeit wurde an keiner anderen Hochschule als Dissertation eingereicht. Ich habe früher noch keinen Promotionsversuch unternommen.

---

Ort, Datum

---

Unterschrift

# List of abbreviations

$\Delta$	Delta
$^{\circ}\text{C}$	Degree Celsius
$\beta$ -2 M	Beta 2 microglobulin
$\mu\text{g}$	Microgram
$\mu\text{l}$	Microliter
$\mu\text{m}$	Micrometer
$\mu\text{M}$	Micromolar
1kGP	1000 Genome Project
BAF	B-allele frequency
BER	Base excision repair
BMP	Bone morphogenic pathway
BF	Bayes factor
bp	Base pair
CCD	Charge-coupled device
cDNA	Complementary deoxyribonucleic acid
chr	Chromosome
cm	Centimeter
CGH	Comparative genomic hybridization
CHRPE	Congenital hypertrophy of the retinal pigment epithelium
CNP	Copy number polymorphism
CNS	Central nervous system
CNV	Copy number variation
CRC	Colorectal cancer
Ct	Threshold cycle
Cyc	Cyclophilin
ddNTP	2',3'-dideoxynucleotide
DEL	Deletion

DEPC	Diethylpyrocarbonate
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
dNTP	Deoxy-nucleotide
DSB	Double strand break
DUP	Duplication
EDTA	Ethylenediaminetetraacetic acid
e.g.	For example
et al.	Et alii (and others)
EtBr	Ethidium bromide
etc.	Et cetera
EtOH	Ethanol
EVS	Exome Variant Server
F	Female
FAP	Familial adenomatous polyposis
FFPE	Formalin fixed paraffin embedded
FISH	Fluorescence <i>in situ</i> hybridization
For	Forward
FoSTeS	Fork stalling and template switching
GI	Gastrointestinal
GO	Gene Ontology
GWAS	Genome-wide association study
HBD	Homozygosity by decent
HI	Haploinsufficiency
HNPCC	Hereditary non-polyposis colorectal cancer
HNR	Heinz-Nixdoff Recall
hr	Hour
kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCR	Low copy repeat

LD	Linkage disequilibrium
LOH	Loss of heterozygosity
LRR	Log R ratio
M	Male
MAF	Minor allele frequency
MALDI-TOF	Matrix-assisted laser desorption/ionization time-of-flight
MAP	<i>MUTYH</i> -associated polyposis
Mb	Megabase
MGB	Minor groove binder
min	Minute
ml	Milliliter
MLPA	Multiplex ligation-dependent probe amplification
mM	Millimolar
mRNA	messenger ribonucleic acid
MS	Mass spectrometer
n/a	Not applicable
NAHR	Non-allelic homologous recombination
ng	Nanogram
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NTC	No template control
OB-HMM	Objective Bayes Hidden Markov Model
OD	Optical density
OMIM	Online Mendelian Inheritance in Man
OR	Olfactory receptor
p	Short arm of chromosome; empirical significance level
PCR	Polymerase chain reaction
pmol	picomol
q	Long arm of chromosome
qPCR	Quantitative polymerase chain reaction

RefSeq	Reference sequence
Rev	Reverse
RNA	Ribonucleic acid
rpm	Revolution per minute
RT	Reverse transcription
RT-PCR	Real-time polymerase chain reaction
SD	Standard deviation
sec	Second
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
T <sub>m</sub>	Melting temperature
TSG	Tumor suppressor gene
U	Unit
UPD	Uniparental disomy
UTR	Untranslated region
UV	Ultraviolet
VNTR	Variable number tandem repeat
WES	Whole exome sequencing
WGS	Whole genome sequencing

# 1. INTRODUCTION

Colorectal cancer (CRC) is one of the most common causes of death in Western countries. In the vast majority of patients, it presents as a sporadic disease. Around 20% of patients exhibit unspecific familial clustering. In up to 5% of the patients, CRC occurs in the context of a monogenic condition caused by highly penetrant germline mutations. The hereditary forms can be divided into two major groups: the more frequent Lynch syndrome or hereditary non-polyposis colorectal cancer (HNPCC), and the various gastrointestinal polyposis syndromes. Relatives of patients with a hereditary type of CRC have a high lifetime risk of gastrointestinal tumors and a syndrome-specific spectrum of extraintestinal malignancies (Aretz 2010). Frequent gastrointestinal endoscopic surveillance is advised for patients and persons at risk and has improved the prognosis considerably (van der Meulen-de Jong et al. 2011).

Colorectal adenomatous polyposis are cancer predisposition syndromes, characterized by the occurrence of dozens to thousands of adenomatous polyps, which, if not detected early and removed, invariably result in CRC. So far, two different inherited forms can be delineated by molecular genetic analysis: the autosomal dominant Familial Adenomatous Polyposis (FAP) caused by heterozygous germline mutations in the tumor suppressor gene (TSG) *APC* on chromosome 5q22, and the autosomal recessive *MUTYH*-Associated Polyposis (MAP) caused by biallelic germline mutations of the base excision repair (BER) gene *MUTYH* on chromosome 1p34 (Aretz et al. 2013; Aretz et al. 2006). In FAP, the vast majority of mutations are truncating point mutations and large deletions ([www.lovd.nl/APC](http://www.lovd.nl/APC)) while the mutation spectrum in MAP is dominated by missense mutations ([www.lovd.nl/MUTYH](http://www.lovd.nl/MUTYH)).

Nowadays, conventional methods including Sanger sequencing and deletion/duplication analysis by MLPA are widely used for the detection of germline mutations of the *APC* and *MUTYH* genes in adenomatous polyposis patients. However, in up to 50% of patients, *APC* and *MUTYH* mutations cannot be identified by these methods although a genetic cause is likely. A reason for this failure might be that mutations in these genes are overlooked by routine molecular methods. However, even if these patients do not carry a germline mutation in known genes, the presence of dozens or more colorectal adenomas argues in favor of an underlying genetic predisposition. Following the discovery of *MUTYH* in 2002, which has become a known gene for up to 25% of *APC* mutation-negative patients with adenomatous polyposis, it is reasonable to assume that there might be yet unidentified causative genes awaiting discovery. Identification of new genetic causes of as yet unexplained disease is important. Once the causative genes are known, we could further understand the pathophysiology of the disease and improve medical care of the families regarding differential diagnosis, estimating the recurrence risks, and predictive testing of persons at risk

Germline copy number variants (CNVs) have been recognised as an important form of structural genetic variation, which also predisposes to human disease (Feuk et al. 2006; Krepischi et al. 2012b). Genome-wide CNV analysis does not require a priori hypothesis about the pathophysiological properties of the responsible genes and can thus be applied to those cases where other methods for identifying causal genes such as linkage analysis or candidate gene approaches are not feasible or promising. During recent years, genome-wide copy number analysis has been used to uncover new predisposing genes in various inherited conditions including familial cancer syndromes (Chen et al. 2013; Lucito et al. 2007; Venkatachalam et al. 2011). It is reasonable to ask whether CNVs, in particular heterozygous deletion CNVs, might also be part of the mutation spectrum in yet unidentified genes underlying unexplained adenomatous polyposis syndromes

Next generation sequencing (NGS) refers to high throughput sequencing technologies. The NGS allows faster and more comprehensive diagnostics in genetically heterogeneous conditions. It can be used to sequence whole genome, whole exome, as well as specific genes of interest to discover novel causing disease mutations (Gilissen et al. 2012).

This thesis was intended to identify novel causative genes responsible for monogenic adenomatous polyposis syndromes by the application of a SNP array based CNV analysis and a targeted Next Generation Sequencing (NGS) approach to a large and well-characterized cohort of unrelated patients with etiologically unexplained colorectal adenomatous polyposis.

## 2. BASIC PRINCIPLES

### 2.1. Hereditary Colorectal Cancer

Colorectal cancer (CRC) is one of most common forms of cancer and a major cause of cancer-related death in the world. The etiological factors and pathogenic mechanisms underlying CRC development appear to be complex and heterogeneous. CRC often begins as a polyp on the surface of the colon mucosa. Most polyps remain benign, but some have the potential to turn cancerous, in particular adenomatous polyps (adenomas). The removal of colon polyps at the time of colonoscopy considerably reduces the subsequent risk of colon cancer (Tuohy et al. 2010).

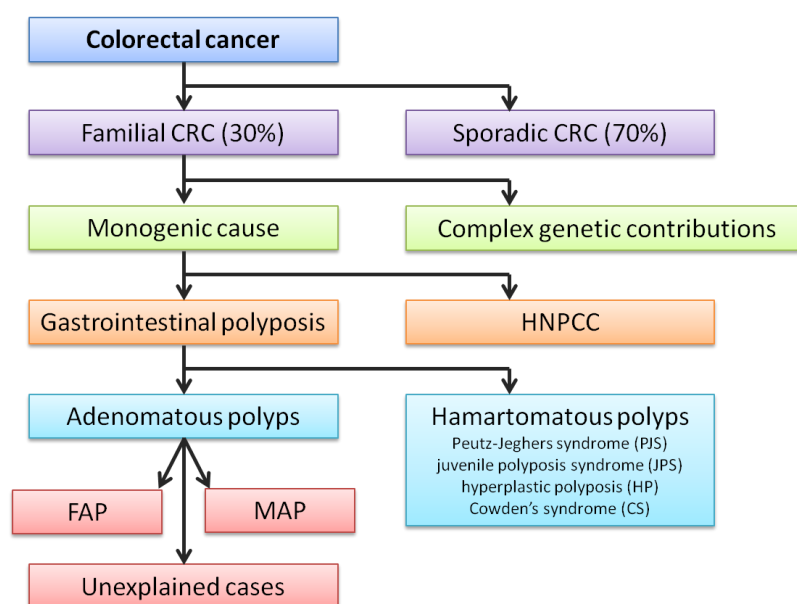
Depending upon the genetics and the etiology of the disease, CRC is usually categorized into sporadic, familial, and hereditary CRC. Sporadic CRC occurs in about 70% of patients, with no apparent evidence of the disorder being inherited. Up to 30% of CRC patients have a positive family history, which suggests a genetic contribution, but less than 6% of them show the trait in a monogenic (hereditary) form (Fearon 2011). Two major hereditary syndromes known so far are hereditary nonpolyposis colorectal cancer (HNPCC)/Lynch syndrome and the various gastrointestinal polyposis syndromes. These two forms of hereditary cases are found only in 3-5% of all CRC (Daley 2010), however, they are also characterized by a broad syndrome-specific extracolonic tumor spectrum.

Lynch syndrome is an autosomal dominant inherited cancer predisposition syndrome caused by germline mutations of mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*) or the *EPCAM* gene upstream of *MSH2*. The majority of germline mutations have been identified in the *MLH1* gene on chromosome 3p21 and the *MSH2* gene on chromosome 2p16. Mutations in any of these genes prevent the proper repair of DNA replication errors. As abnormal cells continue to divide, the accumulated errors can lead to uncontrolled cell growth and possibly cancer. The *EPCAM* gene is not a MMR gene itself but specific 3' deletions of *EPCAM* lead to inactivation of the downstream *MSH2* gene (Kuiper et al. 2011).

The gastrointestinal polyposes, which account for about 1% of all CRC cases, can be categorized histologically into adenomatous polyposis and hamartomatous polyposes (Figure 2.1). The latter include Peutz-Jeghers syndrome, familial juvenile polyposis syndrome, hyperplastic polyposis, and Cowden's syndrome. The two main hereditary adenomatous polyposis syndromes are familial adenomatous polyposis (FAP) and *MUTYH*-associated polyposis (MAP). FAP, the best known and most frequent inherited form of gastrointestinal polyposis, is an autosomal dominant syndrome, caused by germline mutations in the *APC* gene. The classical (typical) form is characterized by the development of hundreds of



adenomas in the second decade of life. MAP is an autosomal recessive inheritance, described for the first time in 2002, and caused by biallelic germline mutations in the base-excision repair (BER) gene *MUTYH*. It is responsible for about 20% of adenomatous polyposis cases without an *APC* mutation (Aretz et al. 2006; Mongin et al. 2012), or up to 0.3% of CRC patients (Nieuwenhuis et al. 2012). Although *APC* and *MUTYH* are the major causes of colorectal adenomatous polyposis, in up to 50% of patients no germline mutations in these two genes can be detected.



**Figure 2.1.** Genetic etiology of colorectal cancer

### 2.1.1. Differential diagnosis

The different types of hereditary CRC can be distinguished from each other by clinical and histopathologic findings, the mode of inheritance, and molecular genetic analysis. To diagnose an adenomatous polyposis, the presence of at least 10-20 synchronous colorectal adenomas is required. Adenomatous polyposis can be classified as FAP if an *APC* germline mutation is identified. The larger the number of adenomas found the greater the likelihood to identify an *APC*-related FAP (Jasperson et al. 2010). Genetic screening for *MUTYH* mutations is usually done if no *APC* mutation is identified as MAP is suspected in those cases as the most important differential diagnosis. In those cases the family history usually does not show an autosomal dominant pattern. Hamartomatous polyps represent developmental malformations that affect the epithelial glands and the underlying lamina propria, where the mucosal components are arranged abnormally but the epithelial cells are not dysplastic. Patients manifest numerous nonadenomatous lesions. The conditions for differential diagnosis are summarized in table 2.1.

**Table 2.1.** Differential diagnosis of inherited CRC

Syndrome	Clinical features	Gene defects
HNPCC	CRC without polyps; other cancers include endometrial, ovarian, and stomach cancer; occasionally brain tumors	<i>MSH2</i> , <i>MLH1</i> , <i>PMS2</i> , <i>MSH6</i>
FAP	Multiple adenomatous polyps (> 100) and carcinomas of colorectum; duodenal polyps and carcinomas; fundic gland polyps in the stomach; congenital hypertrophy of retinal pigment epithelium	<i>APC</i>
Attenuated FAP	< 100 polyps and/or late onset manifestation	<i>APC</i> (predominantly 5' and 3' mutations)
Gardner syndrome	phenotypic variant of FAP, desmoids tumors and mandibular osteomas	<i>APC</i>
Turcot's syndrome	phenotypic variant of FAP, colorectal polyposis brain tumors (mostly medulloblastomas)	<i>APC</i>
<i>MUTYH</i> -associated polyposis	Multiple adenomatous GI polyps	<i>MUTYH</i> (autosomal recessive)
Peutz-Jeghers syndrome	Hamartomatous polyps throughout GI tract and mucocutaneous pigmentation	<i>STK11</i> , <i>LKB1</i>
Cowden's syndrome	Multiple hamartomatous polyps in GI tract and various extraintestinal lesions (breast cancer, endometrial cancer, thyroid cancer)	<i>PTEN</i>
Juvenile polyposis syndrome	Multiple juvenile polyps with predominance in colon and stomach	<i>BMPR1A</i> , <i>SMAD4</i>
Hyperplastic polyposis syndrome	≥ 30 hyperplastic polyposis throughout colon	Genetic cause unknown

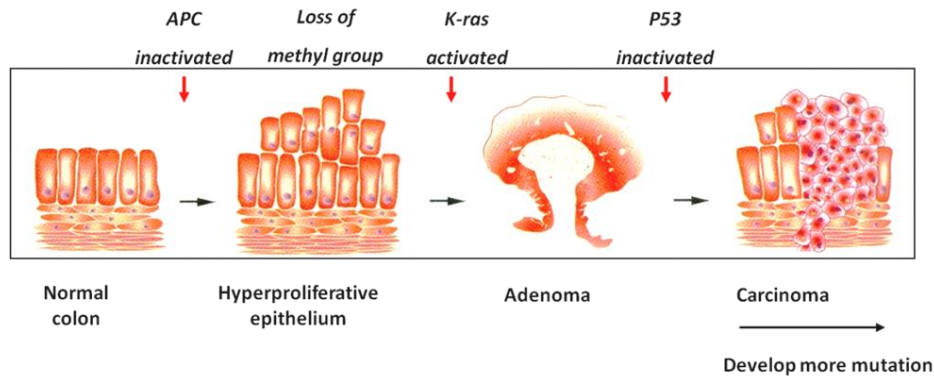
CRC, colorectal cancer; FAP, familial adenomatous polyposis; GI, gastrointestinal; HNPCC, hereditary nonpolyposis colorectal cancer.

### 2.1.2. Adenoma-carcinoma sequence

The majority of CRCs develop over more than 10 years from benign pre-neoplastic lesions. Adenomas are the most common form of premalignant precursor lesions. A multistep model of carcinogenesis for the development of CRC has been described by Vogelstein et al. (1988), known as adenoma-carcinoma sequence. Adenomas arise as the result of the accumulation of genetic and epigenetic changes that transform normal glandular epithelial cells into benign neoplasia, followed by invasive carcinoma and eventually metastatic cancer.

Mutations in a number of oncogenic driver genes (tumor suppressor genes, oncogenes, DNA repair genes) have been implicated in the development of CRC. Mutations in the *APC* gene appear to be one of the earliest events in colorectal tumorigenesis. Beyond the defects in the Wnt-*APC*-beta-catenin signaling pathway (see section 2.2.1), other mutations must occur for the cell to become cancerous. Several major molecular abnormalities may be induced on transition from normal epithelium to carcinoma (Figure 2.2). One important driver is the *TP53* gene, located on chromosome 17p, which normally monitors cell division and kills cells if they

have severe DNA defects (apoptosis). Mutations of the *TP53* gene are commonly seen in CRC (Vogelstein et al. 1988; Vogelstein et al. 2013).



**Figure 2.2.** Adenoma-carcinoma development: an early initiating mutation occurs in the *APC* gene, followed by activation of oncogenes such as *KRAS*, leading to adenomatous polyp formation. Together with methylation alterations, some adenomas progress and become an adenocarcinoma. Mutations of *TP53* and deletions of chromosome 18q are commonly found in carcinomas.

### 2.1.3. Knudson two-hit hypothesis

In 1971, Alfred Knudson proposed the two-hit hypothesis to explain the early onset of an inherited form of retinoblastoma at multiple sites. Although one allele is mutated in the germline, the other, normal, allele is still sufficient to protect against tumorigenesis. If a second hit (mutation) to the wildtype allele copy occurs somatically, the gene is completely inactivated, so that cancer can develop (Knudson 2001). The chance for a germline mutation carrier to obtain a second somatic mutation is much greater than the chances for a non-carrier to get two somatic hits in the same cell. This hypothesis serves as the basis for the understanding of how mutations of tumor suppressor genes (TSG), which usually cause autosomal dominant inherited cancer syndromes, drive cancer.

However, the two-hit model of the *APC* gene that initiates colorectal tumorigenesis does not necessarily result in complete loss of function. Mutant *APC* proteins probably retain some function and the two hits are co-selected to produce an optimal level of Wnt activation. In addition, a three-hit hypothesis has been introduced by Segditsas et al. (2009). They showed that a heterozygous deletion of *APC* represents an effective third hit in cases where the germline mutation is located at the very 5' end of the gene.

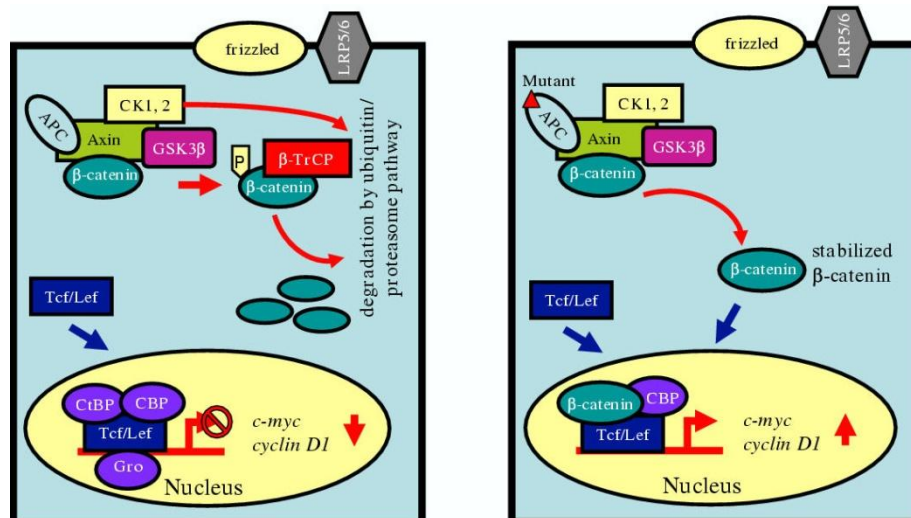
## 2.2. Genetics of adenomatous polyposis syndromes

### 2.2.1. Familial adenomatous polyposis (FAP)

Familial Adenomatous Polyposis (FAP) (MIM# 175100) is an autosomal dominant cancer predisposition syndrome, characterized by the development of hundreds of adenomatous polyps of the colorectum and caused by high-penetrant germline mutations in the *Adenomatous Polyposis Coli (APC)* gene (Groden et al. 1991). It is estimated to occur in 1 of 10,000 individuals and in both genders equally. Adenomas usually occur within the second decade (average age of 16 years) and become symptomatic during the third decade of life. If not removed, the likelihood to progress to CRC is high. The average age of colon cancer diagnosis in untreated individuals is 39 years (Petersen et al. 1991). Attenuated FAP (AFAP) is a mild form of FAP with less than hundred colorectal adenomas and/or late onset of adenoma development. Usually, both adenoma formation and CRC occur 10-15 years later compared to classical FAP.

*APC* is a tumor suppressor gene located on chromosome 5q22 and a central player of the Wnt-signaling pathway. It interacts with the adherens junction proteins and  $\beta$ -catenin. Mutations of *APC* cause aberrant accumulation of  $\beta$ -catenin, which moves into the nucleus, then binds T cell factor-4 (Tcf-4), causing increased transcriptional activation of target genes (Figure 2.3). This results in activation of the Wnt-signaling pathway and contributes to tumorigenesis by altering the relative adhesiveness of colonic epithelial cells and misregulating the integrity of cadherin-catenin complexes. *APC* also plays a role in controlling cell cycle by inhibiting the progression of cells from G0/G1 to the S phase, helping to suppress tumorigenesis. Furthermore, *APC* stabilizes microtubules, thus promoting chromosomal stability. Inactivation of *APC* can lead to defects in mitotic spindles and chromosomal mis-segregation, with the resulting aneuploidy leading to CRC (Galiatsatos and Foulkes 2006). Mutant *APC* proteins could alter  $\beta$ -catenin-mediated cell-cell signaling in a dominant-negative effect (Dihlmann et al. 1999). However, it is likely that this dominant-negative effect alone is not sufficient to produce a tumor, and that additional mutations, such as ras-activating mutations, loss of p53 function, or loss of the wild-type allele may be required for tumor formation.

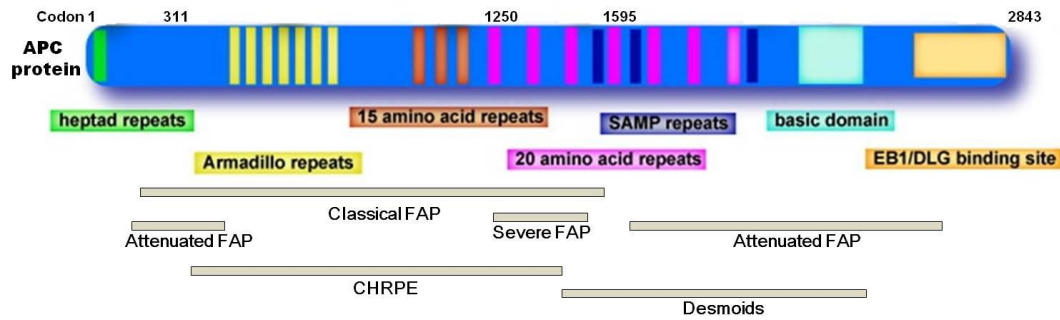
The *APC* gene consists of 15 coding exons and encodes a protein of 2,843 amino acids (Groden et al. 1991; Kinzler et al. 1991). In normal tissue, several isoforms of *APC* messenger RNA (mRNA) are generated as a result of alternative splicing. To date, more than 1,500 different *APC* germline mutations have been identified in FAP patients ([www.hgmd.org](http://www.hgmd.org); [www.lovd.nl/APC](http://www.lovd.nl/APC)). The vast majority is predicted to result in truncated proteins due to nonsense or frameshift mutations, mutations in highly conserved splice sites, or large deletions. Around 60% of mutations occur in exon 15. Two *APC* hot spot mutations are at codons 1061 and 1309 and account for around one third of the identified mutations (Fearon 2011).



**Figure 2.3.** Model of the canonical Wnt signaling pathway. **Left:** The APC protein aligns with other proteins such as glycogen synthase kinase 3 $\beta$  (GSK3 $\beta$ ) to form a destruction complex. This complex phosphorylates  $\beta$ -catenin, leading to degradation of  $\beta$ -catenin. Thus no  $\beta$ -catenin can move into the nucleus. **Right:** Inactivation of APC can lead to the accumulation of  $\beta$ -catenin. When free  $\beta$ -catenin is increased, the  $\beta$ -catenin can enter the nucleus and activate the transcription of TCF-regulated target genes. (Figure adapted from [www.boundless.com/biology/cancer-and-disease/genetic-basis-for-cancer/cancer-development-is-a-multi-step-process/](http://www.boundless.com/biology/cancer-and-disease/genetic-basis-for-cancer/cancer-development-is-a-multi-step-process/))

Differences in phenotypic expression are partly related to the location of the mutation within the *APC* gene (Figure 2.4). Most of the correlations between the mutation site and the clinical phenotype (genotype-phenotype correlation) have been proved to be significant and consistent (Friedl and Aretz 2005). Mutations in the middle part of the gene are associated with 'classical FAP' with early onset of 100-1000 polyps. Severe FAP usually co-occurs with mutations of the *APC* gene between codons 1250-1464, which is a mutation cluster region. The mild form of FAP is usually caused by mutations in the extreme 5' end or the 3' half of *APC* or in the alternatively spliced region of exon 9 (Jones et al. 2002). Mutations in the 3' half of the gene are usually associated with extra-intestinal manifestations such as desmoids. Desmoid tumors generally occur in FAP patients with mutations downstream of codon 1400 (Heinen 2010). The presence of congenital hypertrophy of the retinal pigment epithelium (CHRPE) appears restricted to patients with inherited mutations between codons 311 and 1465.

Two phenotypic variants of *APC*-associated polyposis syndromes are Gardner syndrome and Turcot syndrome (Table 2.1). They can occur in any individual with FAP (Gardner and Richards 1953). Gardner syndrome is characterized by colonic polyposis together with osteomas and soft tissue tumors (desmoids, fibromas). Turcot syndrome is defined as the association of colonic polyposis and central nervous system (CNS) tumors.



**Figure 2.4.** APC mutation spectrum and corresponding clinical outcome (genotype-phenotype correlations). Classical FAP is associated with mutations in codons 157-1595, excluding the mutation cluster region (codon 1250-1464), which is found in FAP patients with severe colorectal phenotype. A milder form of FAP is associated with *APC* mutations in three regions: 1) 5' end of the gene; 2) the alternative splice region in exon 9; and 3) the 3' part of the gene. CHRPE is associated with mutations between codons 311-1465. Desmoid tumors are related to a mutation beyond codon 1400.

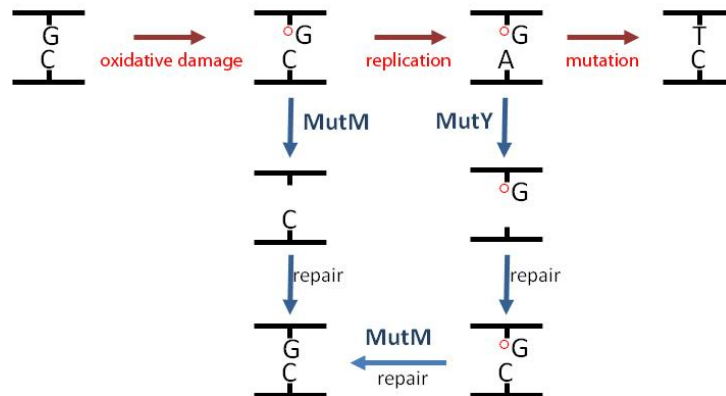
### 2.2.2. *MUTYH*-associated polyposis (MAP)

*MUTYH*-associated polyposis (MAP) (MIM# 608456) is the most important differential diagnosis of FAP. It is the first known polyposis syndrome with a recessive mode of inheritance. MAP is caused by biallelic *MUTYH* mutations and is characterized by ten to a few hundred colorectal adenomas at around 50 years of age. The risk of untreated MAP patients to develop CRC is around 80% at age of 60 years approximately.

*MUTYH* is a base excision repair (BER) gene located on chromosome 1p34.1. Mutations of BER genes lead to an increase of G:C to T:A transversions (Moriya and Grollman 1993). The role of *MUTYH* or the MutY homolog gene in polyposis was discovered in 2002 by Al-Tassan et al. By studying eleven tumors in three affected siblings without inherited mutations of *APC*, they found that 15 out of 18 somatic mutations are G:C to T:A transversions (Al-Tassan et al. 2002). This finding led to the suspicion that the *MUTYH* protein may be deficient in these patients. Many functional studies of this protein as well as mutation screening in many patient cohorts could clarify this hypothesis.

The *MUTYH* gene consists of 16 exons and encodes a protein of 546 amino acids. It encodes a DNA glycosylase enzyme involved in oxidative DNA damage repair. The *MUTYH* glycosylase corrects the intermediate mutational state (incorporation of an A opposite to °G ) during DNA replication before cell division, and thus prevent the accumulation of DNA mutations which might predispose to tumorigenesis. This function is known as BER (Figure 2.5). When BER in the cell is impaired due to *MUTYH* inactivation, mutations in other genes such as *APC* build up, leading to cell overgrowth and possibly tumor formation. To date, more than 300 unique variants of *MUTYH* have been reported in the LOVD database. Most are specific missense mutations, along with small deletions, duplications, and insertions. Two most common missense mutations in the Caucasian population are Tyr179Cys and Gly396Asp. These two mutations have not been identified in Far Eastern Asian populations;

thus, they probably represent founder effects from a common European ancestor (Aretz et al. 2013).



**Figure 2.5.** Base excision repair (BER) pathway; oxidative stress can induce a formation of 8-oxoguanine (8-oxoG) in DNA ( $^{\circ}\text{G}$ ) resulting in a G:C to T:A transversion mutation after replication in the case of an unrepaired 8-oxoG lesion. MutM (OGG1) and MutY (MUTYH) can repair the intermediate mutation states. The MutM (OGG1) will correct the 8-oxoG, and then a repair process will take place. Or if A is misincorporated opposite the 8-oxoG as a consequence of inaccurate replication, MutY (MUTYH) will remove it and resynthesis will be accomplished.

### 2.2.3. Mutation negative adenomatous polyposis

In up to 50% of patients with adenomatous polyposis, no *APC* or *MUTYH* germline mutation can be identified in routine diagnostics. However, the occurrence of dozens or hundreds of colorectal polyps strongly argues for an underlying genetic predisposition. One possibility to explain mutation-negative adenomatous polyposis is the limitation of present routine diagnostics methods, which leads to overlooking mutations in both genes. Aretz et al. (2007a) and Hes et al. (2008) reported mosaic mutations in the *APC* gene in a substantial number of FAP patients with an empty family history, in whom no *APC* mutation was identified by routine screening methods, because the signal of the mutated allele was very low. Moreover, there might be additional mutations in non-scanned parts (non-coding regions) of the genes, which have not been discovered with present methods.

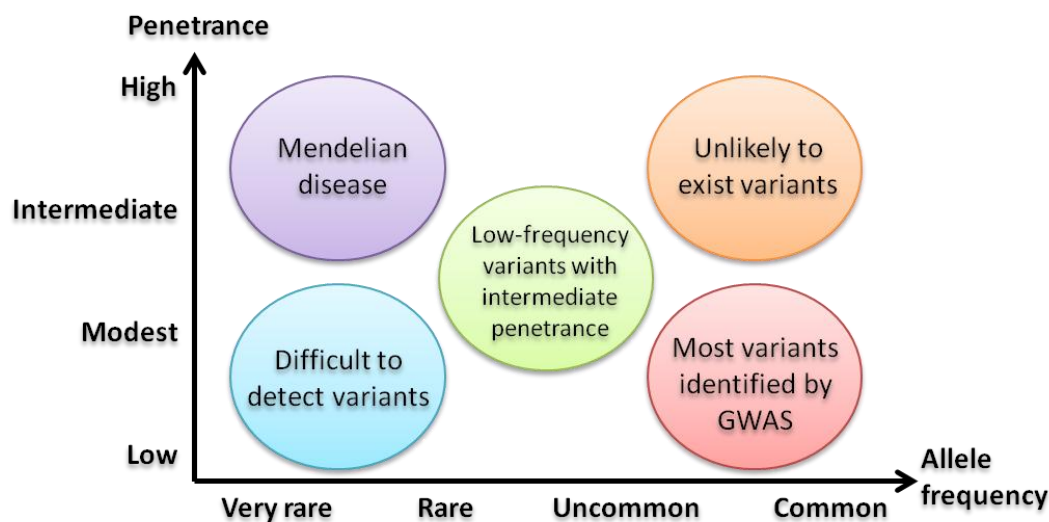
The high number of adenomas and the early age at onset cannot be explained convincingly by non-genetic/environmental factors. Thus, it is likely that a yet unknown inherited predisposition is underlying the unexplained adenomatous polyposis cases, either as a monogenic trait or in a more oligogenic/multifactorial way which means, a fairly large number of mutated genes may contribute to disease, each with only a small effect that is difficult to uncover. As the majority of unexplained adenomatous polyposis patients are sporadic cases, monogenic subtypes are likely to be autosomal recessive.



## 2.3. Human genome variation

The human genome consists of about 3.2 billion base pairs (bp), and it is currently estimated that about 21,000 protein coding genes are encoded in the human DNA (Genome Reference Consortium). Protein-coding sequences make up only about 1-2% of the human genome.

Genomic variations have different forms: 1) Single nucleotide polymorphisms (SNP); 2) tandem repeats; and 3) structural variants including deletions, duplications, translocations, and inversions (Ku et al. 2010). All types of genetic variation can also occur as rare and high penetrant mutations which may cause human disease.



**Figure 2.6.** Graph explaining the expected allele frequency and penetrance of rare and common diseases. The majority of Mendelian diseases are caused by very rare variants with high penetrance (McCarthy et al. 2008).

A common variant is often defined as having a minor allele frequency (MAF) of more than 1% in the general population. Conversely, the definition of rare variant refers to a variant with a MAF of less than 1% (Ionita-Laza and Ottman 2011; Lefevre et al. 2012). Variants found frequently in the human genome and may or may not have an impact on the phenotype are also called 'polymorphisms'. Mendelian (monogenic) diseases are usually caused by a rare variant with high penetrance. However, some mutations observed in autosomal recessive conditions such as Cystic Fibrosis or MAP, have an MAF of  $> 1\%$ . In contrast to Mendelian disorders, multifactorial diseases are caused by several common variants with low penetrance (common disease - common variant hypothesis) (McCarthy et al. 2008). Penetrance is defined as the proportion of mutation carriers that express the related phenotype (Figure 2.6).



### **2.3.1. Single nucleotide polymorphisms (SNPs)**

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation characterized by single-base exchanges at a specific DNA site and are distributed throughout the genome. The human genome contains at least 11 million SNPs that occur with a minor allele frequency (MAF) of at least 1% (International-HapMap-Consortium 2005). They are estimated to occur at a frequency of approximately one per 300 nucleotides. Therefore, for every 300 nucleotides, the average nucleotide identity at one position will differ between any two copies of that chromosome at a substantial frequency throughout a population. A wide variety of approaches for genotyping SNPs have been developed, one approach being Micro-array technology (Carter 2007). SNP genotyping is important in human genetic studies because a specific SNP allele can be implicated as an associated causative factor in human genetic disorders and can be used as genetic marker for genetic-mapping studies.

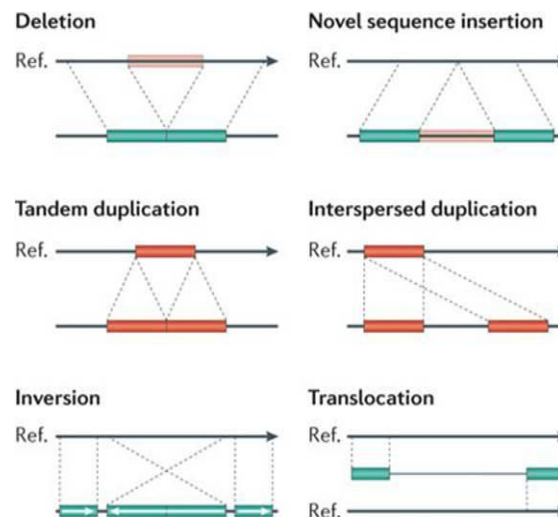
### **2.3.2. Tandem repeats**

Tandem repeats are repeats of a nucleotide pattern and can be divided into two classes, short tandem repeats (STRs) and variable number tandem repeats (VNTRs). STRs or microsatellites are repeats typically composed of 2-7 nucleotides. These markers are abundant, equally distributed throughout the genome and are highly polymorphic compared with other genetic markers. STRs have become popular DNA markers; the number of repeats in STR markers can be highly polymorphic among individuals, which makes STRs effective in the identification of genes underlying monogenic conditions by linkage analysis. VNTRs, also known as minisatellites, consist of repetitive sequences of more than eight bases (commonly 10 – 60 base pairs).

### **2.3.3. Structural variations**

Variants comprising more than a single nucleotide are broadly defined as structural variations. They include insertions, deletions, translocations, duplications and inversions (Figure 2.7). The sizes of structural variants range from the microscopic level, e.g., chromosomal aberrations, to the copy number variant level, which is on the order of  $\geq 1$  kb. One of the smallest structural variants are 'indels', which are defined as either an insertion or deletion of nucleotides up to 50 bases in length (Montgomery et al. 2013). Next to SNPs, indels are the second most abundant form of genetic variation. It has been estimated that there are 1-2 million short indels segregating at low to high frequency in modern human populations. The vast majority of indels occur in short tandem repeats. Many of these indels map to functionally important sites within human genes and are likely to influence human traits and diseases (Mullaney et al. 2010).

Several studies suggest that structural variants account for at least 70% of all variant bases in the human genome, and for any given individual, structural variations constitute between ~0.5% to 1% of the genome (Alkan et al. 2011). All these variations likely contribute to both human diversity and disease susceptibility due to altered gene dosage levels, disruption of proximal or distant regulatory regions, or by affecting the coding sequence.



**Figure 2.7.** The schema represents types of structural variants compared to the reference genome (upper line); deletion, insertion, duplication, inversion, and translocation. (Figure adapted from Alkan et al. (2011)).

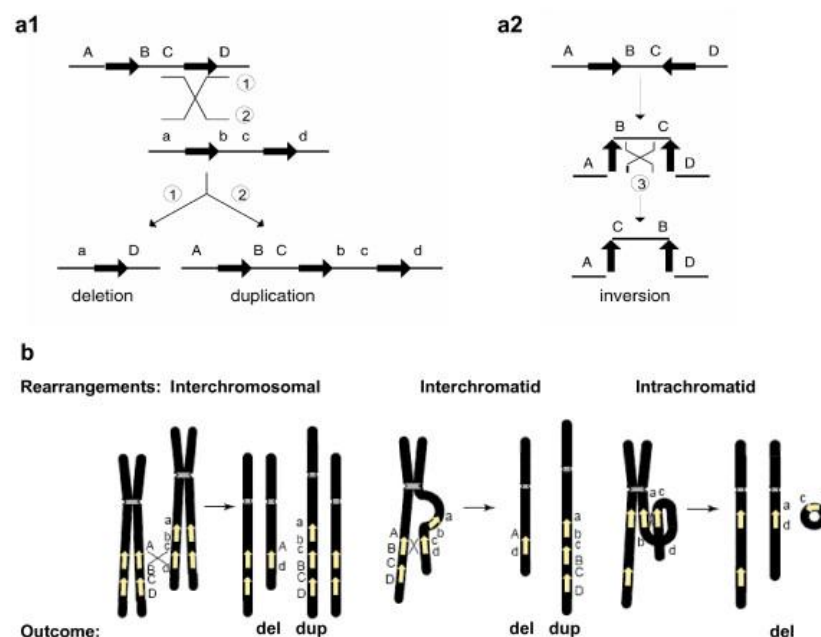
## Copy number variation (CNV)

Copy number variation (CNV) is a major part of human genetic variation (Cook and Scherer 2008) and an important class among the other types of structural variation (Feuk et al. 2006). As humans have two copies of each autosomal chromosome, a deletion or a duplication of one allele alters the number of copies, and this phenomenon is called copy number variation (Freeman et al. 2006). Thus, CNVs include gains (insertions or duplications) and losses (deletions or null alleles) of genomic regions. By definition, the term refers to DNA regions  $\geq 1$  kb in length. The majority of CNVs are 1-10 kb in length.

### ***Mechanisms of CNV formation***

Three major mechanisms have been proposed to cause genomic rearrangements in the human genome: (1) Non-allelic homologous recombination (NAHR) is well known, mostly mediated by low-copy repeats (LCRs) with recombination hotspots, gene conversion, and apparent minimally efficient processing segments; (2) non-homologous end joining (NHEJ); and (3) Fork Stalling and Template Switching (FoSTeS) models, both of which are responsible for non-recurrent rearrangements (Gu et al. 2008).

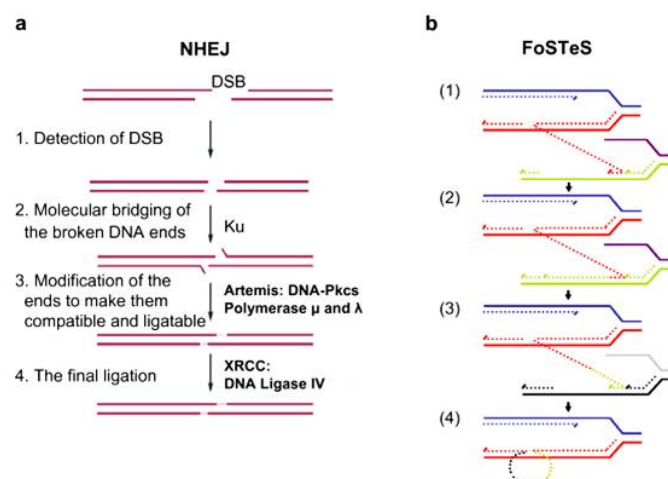
**1 Non-allelic homologous recombination (NAHR)**, also called 'Hotspots inside low-copy repeats', is often caused by LCRs. LCRs are segmental duplications which are defined as regions longer than 10 kb with over ~97% sequence identity. LCRs can cause genomic instability and either mediate or stimulate CNV formation. When LCRs are located at a distance less than 10 MB from each other, they can lead to misalignment in both meiosis and mitosis, and mediate non-allelic homologous recombination (NAHR), which can result in unequal crossing-over. When two LCRs are located on the same chromosome in the same direction, NAHR between them causes a duplication and/or a deletion (Figure 2.8). When they are on the same chromosome but in opposite directions, NAHR results in an inversion of the fragment flanked by them. Prominent examples of disease causing CNVs mediated by LCRs are a number of well known human microdeletion syndromes such as Angelman syndrome, DiGeorge syndrome, or Williams-Beuren syndrome. Non-homologous recombination events that underlie changes in copy number also allow for generation of new combinations of exons between different genes by translocation, insertion, or deletion, so that proteins might acquire new domains, and hence new or modified activities (Hastings et al. 2009).



**Figure 2.8.** Nonallelic homologous recombination (NAHR): **a1** and **a2**) Genomic rearrangements resulting from low-copy repeat (LCR) recombinations. Black arrows represent LCRs and the direction of the arrowhead indicates the orientation of LCRs. Letters indicate the flanking sequences. Thin diagonal lines refer to recombination events with results identified by numbers 1, 2, and 3. **b**) Schematic representation of reciprocal duplications and deletions mediated by interchromosomal (left), interchromatid (middle) and intrachromatid (right) NAHR using LCR pairs in direct orientation. Interchromosomal and interchromatid NAHR between LCRs in direct orientation result in reciprocal duplication and deletion, whereas intrachromatid NAHR only creates deletion (Figure adapted from Gu et al. (2008)).

**2 Non-Homologous End Joining (NHEJ)** is one of the two major mechanisms for repairing DNA double strand breaks (DSB); it is used to explain duplications (Lee et al. 2006). Random breakage can cause large inverted duplications, and repeated cycles could lead to amplification of the inverted repeat. The breakage-fusion-bridge cycle has been linked to the formation of amplification and is believed to play a major role in amplification in cancer. In contrast to NAHR, NHEJ does not require LCRs to mediate the recombination but may also be stimulated by genome architecture. When a DSB is detected, then both broken DNA ends are bridged, modified, and finally ligated (Figure 2.9) (Lieber 2008).

**3 Fork Stalling and Template Switching (FoSTeS)** is the third major mechanism for human genomic rearrangement (Lee et al. 2007) and might be a major mechanism for duplication CNVs. The mechanism is based on DNA replication errors. During DNA replication, the DNA replication fork stalls at one position; one replication fork with a lagging strand invades and anneals to another replication fork in physical proximity at the 3' end and starts the DNA synthesis. The priming results in a 'join point' rather than a breakpoint and can result in complex rearrangements. Depending on the location of the new fork, a deletion or a duplication will occur (Gu et al. 2008).



**Figure 2.9.** Genomic rearrangement mechanisms: **a)** Non-homologous end-joining (NHEJ); a double-stranded DNA break (DSB) occurs and is repaired via NHEJ mechanism. The two thick lines depict two DNA strands with DSB, the thin segments in the middle represent the modifications, through which the ends have gone before the final ligation. Ku is a DNA end-binding protein; once Ku is bound to the DNA end it can improve the binding equilibrium of the nuclease (Artemis-DNA-Pkcs), polymerases ( $\mu$  and  $\lambda$ ), and ligase (XRCC-DNA ligase IV) of NHEJ. **b)** Fork stalling and template switching (FosTes); after the original stalling of the replication fork (dark blue and red, solid lines), the lagging strand (red, dotted line) invades and anneals to a second fork (purple and green, solid lines) followed by DNA synthesis (green, dotted line). After the fork disengages, the original fork (dark blue and red, solid lines) with its lagging strand (red and green, dotted lines) could invade a third fork (gray and black, solid lines). Dotted lines represent newly synthesized DNA. Serial replication fork disengaging and lagging strand invasion could occur several times before resumption of replication on the original template (Figures adapted from Lee et al. (2007) and Gu et al. (2008)).

### ***Functional impact of CNVs and disease association***

Nowadays, large-scale duplications and deletions are thought to be a normal part of genetic variation. Like SNPs, CNVs have become important in genetic diversity as they play a role in genetic susceptibility to common diseases including cancers (Ionita-Laza et al. 2009; Krepischi et al. 2012b; Stankiewicz and Lupski 2010). In hereditary cancer syndromes, rare large heterozygous deletions in several known cancer-predisposing genes such as *APC*, *BRCA1*, *BRCA2*, *RB1* and *TP53* substantially contribute to the germline mutation spectrum (Kuiper et al. 2010). Thus, common and rare CNVs play important pathogenic roles, from causative high penetrant CNVs (mostly deletions) in rare genomic disorders to intermediate or low penetrant CNVs in complex multifactorial diseases (Fanciulli et al. 2010). Consequently and similar to SNPs and rare point mutations, common and rare CNVs should be considered separately as they may play different roles in cancer predisposition (Shlien and Malkin 2009). In line with the definition of SNPs, common CNVs shared by > 1% of the population are also referred to as copy number polymorphisms (CNP), however, the term 'CNP' is not consistently used. CNPs correspond mostly to ancestral events and segregate in the population with different allele frequencies (McCarroll et al. 2008; Redon et al. 2006).

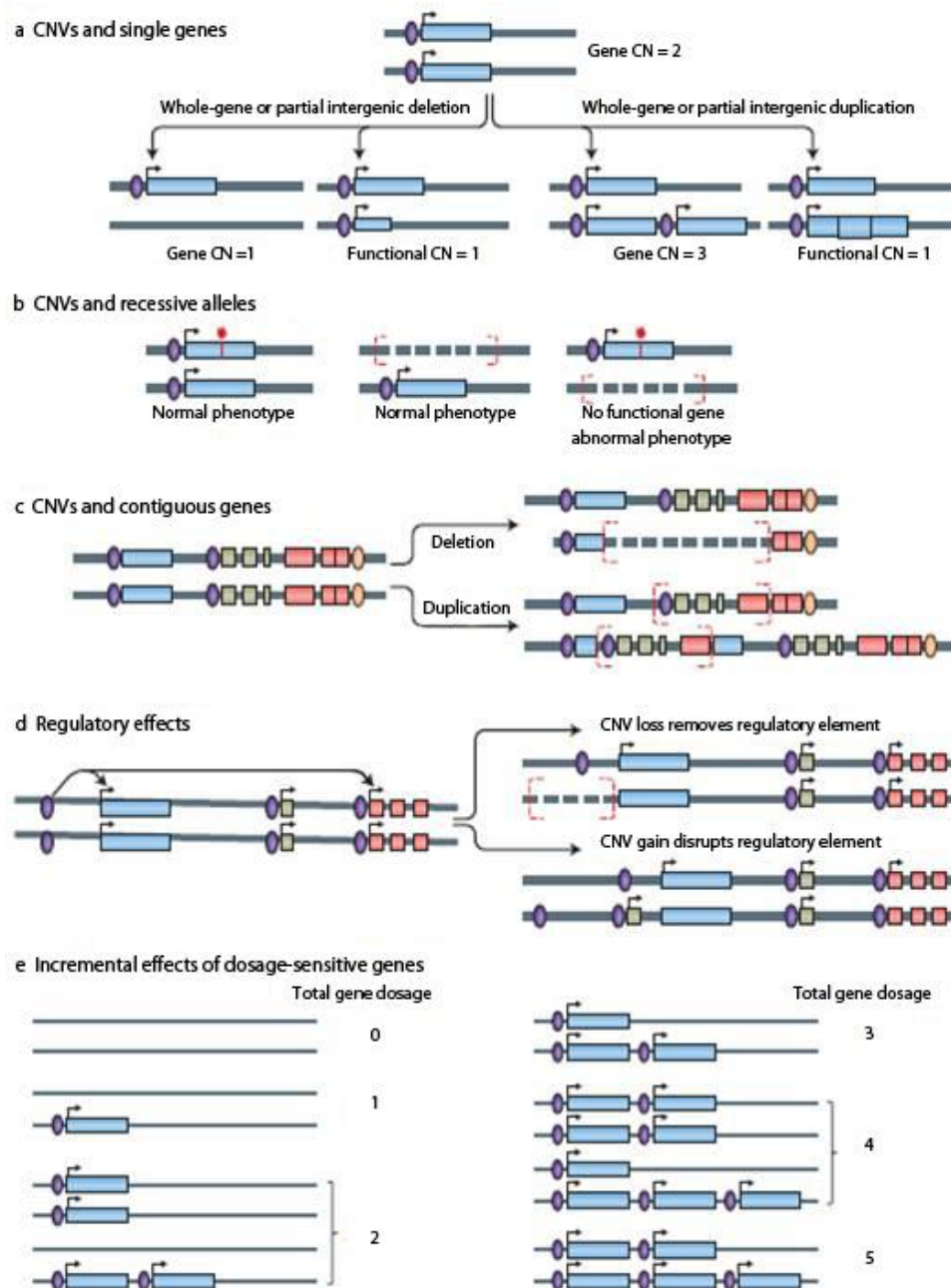
**CNVs and gene expression:** Changes in copy number might change the expression levels of genes included in the affected region, thus allowing transcription levels to be higher or lower (Hastings et al. 2009). By gene expression analysis, around 76% of genes affected by copy number alteration had expression patterns that correlated with the gene copy number (Fanciulli et al. 2010; Henrichsen et al. 2009; McCarroll et al. 2006; Orozco et al. 2009; Stranger et al. 2007).

Partial gene duplications might be more deleterious than full gene duplications as they may introduce a frameshift into the transcript or produce a dominant negative isoform when they are ordered in a tandem position, either of which may decrease protein function. Likewise, a chromosomal rearrangement resulting from CNVs can cause a 'positional effect', i.e., the absence of the natural regulatory region or the conjunction with a different regulatory element can change the expression of the respective gene (Kleinjan and van Heyningen 1998).

Stranger et al. (2007) reported a gene expression study in HapMap lymphoblast cells showing that half the effects of known CNVs are caused by gene disruption or by affecting regulatory or other functional regions, not by altering gene dosage. For example, a deletion that moves an enhancer within functional distance of a gene can increase expression of that gene, while a duplication that moves an enhancer further away may decrease gene expression (Figure 2.10). These effects lead to difficulties in predicting functional consequences of CNVs.

Notably, 53% of genes whose expression was influenced by CNVs had the corresponding CNV outside of the actual gene, suggesting that many CNVs could affect important regulatory sequences that are situated at a distance from the actual target gene (Ionita-Laza

et al. 2009). Large *de novo* CNVs, which encompass many genes and/or regulatory sequences, are thought to likely be disease causative (de Smith et al. 2008).



**Figure 2.10.** Illustration of mechanisms of CNVs leading to malfunction of genes. Squares represent genes, ovals represent promoters, and bent arrows represent the direction of transcription (Figure adapted from Lee and Scherer (2010)).

In general, large deletions are supposed to cause more severe phenotypes compared to large duplications since the loss of a copy has more dramatic effects than the gain of a copy in dosage sensitive genes. CNVs that are intragenic or involve a single exon may have functional consequences that are similar to point mutations, behaving much like classical Mendelian dominant or recessive traits. In addition, a deletion may result in

haploinsufficiency for a dosage-sensitive gene. A duplication may create imbalances due to excess product of the duplicated genes, or, when intragenic, may alter the structure of a product and thereby its function (Lee and Scherer 2010). Alternatively, CNVs that overlap genes can result in fusion genes that may have phenotypic consequences.

**CNVs and cancer:** Rare germline CNVs such as deletions and amplifications can cause monogenic genetic disorders including hereditary cancer syndromes (Lucito et al. 2007). So far, more than 30% of approximately 100 known cancer genes predisposing to hereditary tumor syndromes have been observed to include deleterious CNVs; most of them are heterozygous deletions, which disrupt single exons up to the whole gene (Krepischi et al. 2012b; Kuiper et al. 2010). Common germline CNVs are supposed to lead to low penetrance cancer predisposition. Shlien and Malkin (2010) reported 49 cancer-related genes involving common CNVs, most of which presumably have low penetrance and exert only a small contribution to cancer risk. Although common low penetrant CNVs are only modest contributors to cancer individually, their combined impact on cancer predisposition must be taken into account in estimating cancer risks.

**CNVs and colorectal cancer:** Although most highly penetrant genes are frequently affected by point mutations, CNVs, in particular deletions, contribute significantly to the mutation spectrum of known genes underlying hereditary CRC syndromes: in FAP, Lynch syndrome, or Peutz-Jeghers syndrome, around 5-30% of the underlying germline mutations represent heterozygous deletions (Aretz et al. 2005; Krepischi et al. 2012b; Kuiper et al. 2010; Lucci-Cordisco et al. 2005; Shlien and Malkin 2009). Monogenic cancer syndromes such as Lynch syndrome can also be caused by germline CNVs outside the relevant genes such as micro-deletions of the *EPCAM* gene upstream of *MSH2* which causes a transcriptional read-through and epigenetic silencing of *MSH2* (Kuiper et al. 2010).

Although CNVs are a significant submicroscopic form of genetic variation, their influence on phenotypic variability, including disease susceptibility, remains incompletely understood and needs to be investigated further.

## 2.4. Identification of causative genes in human disease

It is important to identify the genes implicated in hereditary diseases, since this knowledge can lead to improvements in differential diagnosis, estimating recurrence risks, disease prevention (surveillance), and treatment. To localize disease predisposing genes, different methods can be applied. The usual process to identify highly penetrant causative genes is linkage analysis in DNA samples of family members affected and not affected with the disease. This approach can only be applied if multiple affected family members are present and available and when the disease status of the family members can be clearly assessed. Homozygosity mapping is another traditional approach to identify the causative gene for



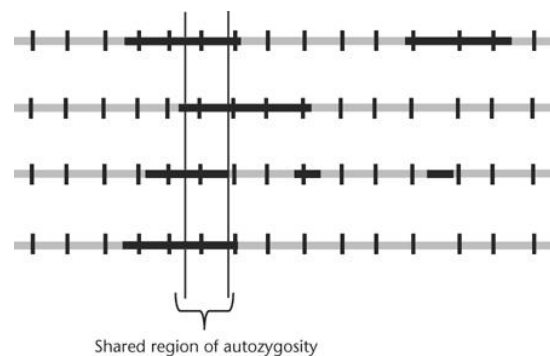
autosomal recessive conditions in consanguineous families. Since the completion of the Human Genome Project, family-based analysis has shifted to hypothesis-free research including genome-wide CNV analysis. Most recently, high-throughput sequencing technologies, in particular exome and genome sequencing, are increasingly used to identify pathogenic mutations in new genes underlying Mendelian disorders (Gilissen et al. 2012; Goldstein et al. 2013).

In multifactorial/polygenic (complex) diseases, genome-wide linkage analysis and SNP-based association studies of candidate genes were used to uncover the genetic basis. However, these methods were replaced recently by genome-wide association study (GWAS) and next generation sequencing (NGS) studies to identify common low penetrant and rare moderate penetrant variants, which constitute the genomic architecture of genomic predisposition.

Below, the most common methods for localizing or directly identifying disease-predisposing genes are outlined in more detail.

### 2.4.1. Homozygosity mapping

Homozygosity mapping is an important tool to identify the causative gene in autosomal recessive conditions in consanguineous families. In most situations, the power for the detection of recessive genes is lower compared to autosomal dominant genes because recessive diseases usually are found only in a single sibship, often with just sporadic appearance, while dominant traits tend to occur in many generations of larger family pedigrees.



**Figure 2.11.** Illustration of the basic principles of homozygosity mapping. The genomic region is represented by a straight grey line with markers indicated by small bars. Regions of autozygosity (homozygous with two identical-by-descent alleles) are in black. Considering four patients who are affected by the same recessive disorder, because of inbreeding, each patient might have several autozygous regions over the genome. However, they share a region of autozygosity, where the disease locus is likely to map. Patients will all be homozygous at the markers located in this region. (Figure adapted from [www.els.net/WileyCDA/ElsArticle/refId-a0005407.html](http://www.els.net/WileyCDA/ElsArticle/refId-a0005407.html))



The basic concept of homozygosity mapping is to trace the inheritance of the same chromosomal region from an ancestor via two consanguineous heterozygous parents and hence homozygosity in the patients; thus, the disease region must be homozygous in all affected family members (Figure 2.11). However, detection of identical regions and homozygosity by descent (HBD) when family data are not available, or when relationships are unknown, is still a challenge (Zhang et al. 2011). Making use of population data from high-density SNP genotyping may allow detection of regions HBD from recent common founders in singleton patients without genealogy information. However, the homozygous regions may include dozens or hundreds of candidate genes. If the affected family members are more distantly related, the homozygous regions will be reduced regarding number and size.

#### **2.4.2. Loss of heterozygosity (LOH) analysis**

A region derived from each parent, which contains different alleles of a genetic variant, is said to be heterozygous. However, one parental copy of a region can sometimes be lost due to non-disjunction during mitosis, segregation during recombination, or deletion of a chromosome segment, which results in the region having just one copy. That single copy cannot be heterozygous and therefore the region shows a “loss of heterozygosity” (LOH). To demonstrate heterozygosity, genetic markers such as microsatellites or SNPs can be used.

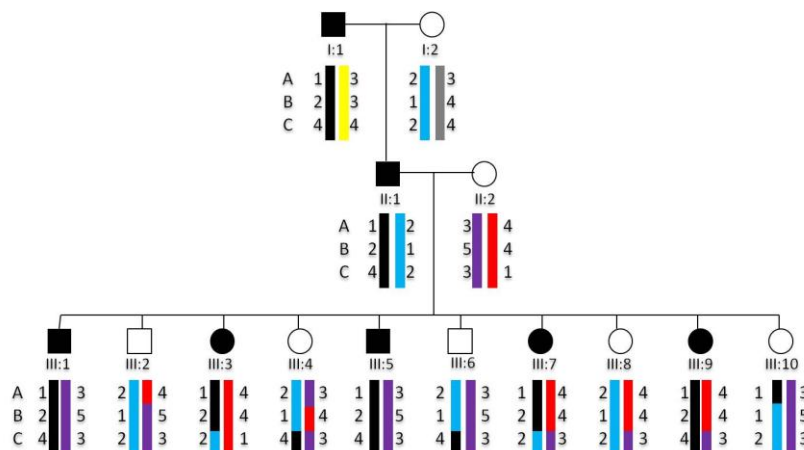
LOH becomes critical when the remaining allele contains a point mutation that renders the gene inactive. This alteration is a common genetic occurrence in cancers where a tumor suppressor gene (TSG) is affected. The LOH of a TSG, often examined by genotyping microsatellite markers flanking the gene in normal tissue compared to tumor tissue, is an important marker for loss of the second (wildtype) allele in tumor tissue. The deletion of the wildtype allele is the second hit required for TSG inactivation and is believed to be one of the key steps in the carcinogenesis of tumors such as CRC. Analysis of LOH is important in cancer research for localizing potential TSGs (Paulson et al. 1999). There have been several reports of LOH in candidate TSGs in CRC, which relate to chromosomes 7, 8, 18, 20, and 22 (Eldai et al. 2013; Therkildsen et al. 2013; Yam et al. 2013).

Microsatellite markers, also known as short tandem repeats (STRs), are polymorphic DNA loci consisting of 2-6 repeated nucleotide sequences. The number of repeat units varies in the population. LOH using microsatellite markers is used for screening of tumor samples by comparing suspected cancerous tissue and healthy tissue from the same individual. Another approach for analyzing LOH is SNP array genotyping to compare SNP genotypes between tumor DNA samples and patient-matched germline DNA samples obtained from normal (non-tumor) tissue. SNP array genotyping data is also useful for the delineation of minimally lost regions that indicate the presence of important TSGs.

With SNP array genotyping data, it is possible to identify copy-neutral LOH. The LOH is called copy-neutral because no net change in the copy number occurs in the affected individual. Other names for copy-neutral LOH are acquired uniparental disomy (UPD) or gene conversion. In UPD, a person receives two copies of a chromosome, or part of a chromosome, from one parent and no copies from the other parent due to errors in meiosis I or meiosis II.

### 2.4.3. Linkage analysis

Linkage is the tendency for genes and other genetic markers to be inherited together because of their location near one another on the same chromosome. The biological basis for linkage analysis is recombination caused by crossing-over during meiosis. The smaller the physical distance between two genetic loci on a chromosome, the lower the frequency of recombination is between the markers, which means that the probability that they are separated by recombination is low. In contrast, the recombination rate is higher (no linkage) if two loci are at opposite ends on a chromosome (Goldstein et al. 2001). Linkage analysis is a traditional approach to search for a disease gene, used to investigate in a family with several affected persons to see if some of a set of markers (microsatellites, SNPs) co-segregate with the phenotype. The co-segregating markers are assumed to be close to a neighboring disease causing gene (Figure 2.12). Linkage studies require no prior information regarding the causative gene, however, the results are influenced by the distance of the markers to the causative region (recombination events), the ability to clearly recognize the phenotype of a proband, and the correctness of the reported relationships between family members.



**Figure 2.12.** The co-segregation of an autosomal dominant disease trait and alleles at 3 polymorphic marker loci (A, B, C). The disease is transferred from I:1 to II:1 by the black allele which contains the haplotype 1-2-4 of the markers A, B, and C. Recombination events are shown in the third generation. Marker C shows recombination in III:4, III:6 with no phenotype and marker A shows one recombination in the unaffected individual III:10. The affected III:3, III:7 demonstrate that the disease trait shows linkage to markers A and B, and the causative gene is located within haplotype block B. (Figure adapted from Pulst (1999)).

Genome-wide linkage analysis is a powerful tool for localizing genes for diseases following Mendelian patterns in families (Goldstein et al. 2001). Microsatellite markers are used as tools for tracking the not yet identified gene. Approximately 400-500 microsatellite markers distributed over the whole genome are examined for genetic mapping. They are based on the observation that genes that reside physically close on a chromosome remain linked during meiosis but the regions identified are often large and include many candidate genes. After the identification of a putative locus, the region that could possibly contain a causal variant is smaller and easier to systematically study.

#### **2.4.4. Genome-wide association study (GWAS)**

Genome-wide association studies (GWAS) aim to identify common genetic variants that might underlie multifactorial, genetically complex disease where a combination of various low or moderate penetrance susceptibility alleles and environmental factors cause the phenotype. A GWAS requires no prior information regarding potential causative genes but is based on the “common disease – common variant” hypothesis. Depending on the type of control individuals/genotypes used, two different approaches are distinguished, family-based or case-control association methods (Ioannidis et al. 2010). In GWAS, to achieve comprehensive coverage and adequate statistical power to detect unknown disease variants through linkage disequilibrium (LD), a large number of genetic markers are required spanning the whole genome as well as a large number of patients and controls.

In an association analysis, unrelated case and control individuals are compared regarding the frequencies of alleles or genotypes of a single marker. Fortunately, the existence of LD significantly reduces the number of SNPs to a set of representative tagSNPs that needs to be genotyped in a GWAS. Genomic regions of interest may also be investigated by haplotype analysis, in which a handful of alleles transmitted together on the same chromosome are tested for association with disease; in this case, the loci which are jointly considered are located within a small genomic region, often confined to the neighborhood of a single gene (Braun and Buetow 2011).

Genetic variants which are identified to be statistically associated with a disease may directly increase susceptibility to a condition, or the associated marker alleles may be linked to nearby causative alleles. More often than not, the markers do not themselves play a role but instead are in LD to the real causative markers. Additionally, they might modify the expression of a nearby gene as most of them are located in regulatory regions. Significant frequency differences of SNPs between patients and controls are interpreted as being due to an original disease mutation having occurred generations ago and a paucity of recombination between disease and neighboring marker loci. Most markers used in GWAS have no effect on the amino acid sequence of a protein since they are silent variants or are located in intergenic or intronic regions with no obvious connection with protein expression.

In recent years the use of genome-wide SNP microarrays (SNP arrays) has exploded due to the rise of large numbers of GWAS. This strategy has been very successful in identifying new genetic loci for various human complex traits (see updated GWAS catalog: [www.ebi.ac.uk/fgpt/gwas/](http://www.ebi.ac.uk/fgpt/gwas/)). Most of the genes and loci identified have not previously been thought to be associated with their respective diseases.

#### 2.4.5. Copy number variation (CNV) analysis

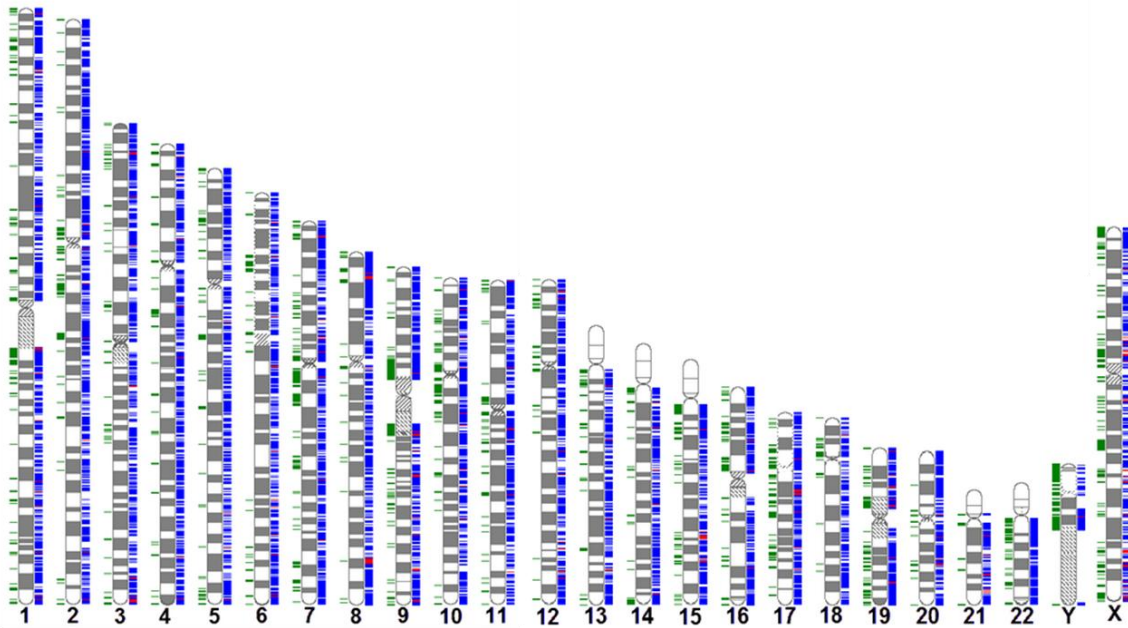
Copy number variation (CNV) is an important form of structural variation. Redon et al. (2006) published the first CNV map, which collected 1447 copy number variable regions, covering 360 Mb (12%) of the human genome. The CNV regions cover more nucleotide content per genome than SNPs; 10% of known genes contain CNVs. Since then, several studies of CNVs with high-throughput technologies have been performed to increase the knowledge of CNV distributions, sizes, frequencies, and other population-genetic parameters (Conrad et al. 2010; Cooper et al. 2008; Jakobsson et al. 2008; Mills et al. 2011). The number of CNVs reported in the *Database of Genomic Variants* (DGV, <http://projects.tcag.ca/variation/>), a catalog of human genomic structural variation, is increasing continuously. In 2008, Perry et al. (2008) reported that CNV regions have been estimated to cover 18% of the human genome. One year later, Zhang et al. (2009) showed that CNVs cover 29.7% of the human genome. In June of 2013, DGV reported a number of 184,148 CNVs by including 53 studies. CNVs are found in every chromosome including sex chromosomes (Figure 2.13).

##### **Detection of CNVs**

Traditional approaches to identify CNVs (large deletions and duplications) employ chromosome analysis and targeted approaches such as karyotyping and fluorescence *in situ* hybridization (FISH), segregation analysis of polymorphic marker alleles, quantitative PCR, RNA analysis, multiplex ligation-dependent probe amplification (MLPA). Nowadays, high-throughput technologies enabling genome-wide discovery of CNVs are broadly used. They can be divided into two categories: hybridization-based and sequence-based approaches (see section 2.4.6). In large-scale genetic studies, hybridization-based technology is the primary method for CNV detection (Pinto et al. 2011). Several genome-wide studies of copy-number variation using hybridization-based methods have been published (Conrad et al. 2006; Cooper et al. 2008; Pinto et al. 2007; Redon et al. 2006). The two main types of microarrays are comparative genomic hybridization (CGH) arrays and single nucleotide polymorphism (SNP) arrays, also called genotyping arrays.

In a typical array-CGH experiment, total genomic DNAs are extracted from patients and references and differentially labeled with two fluorescent dyes, green and red. The two sets of DNA probes are hybridized together to an array (Figure 2.14A). Fluorescence signals are measured separately for the test and the reference samples with two different color channels. For a given probe, the relative intensity of the test versus reference signals reflects a noisy

measurement of the relative DNA amount (Pinkel and Albertson 2005). With proper normalization, the data usually take the form of log ratio (LogR) of test and reference intensities at each probe, linearly ordered according to the physical locations of probes along the genome. The generated CNV profile falls into a wide range of coverage and resolution. Although array-CGH can detect deletions or duplications in very small segments of chromosomes, it cannot detect copy number neutral differences associated with loss of heterozygosity (LOH).



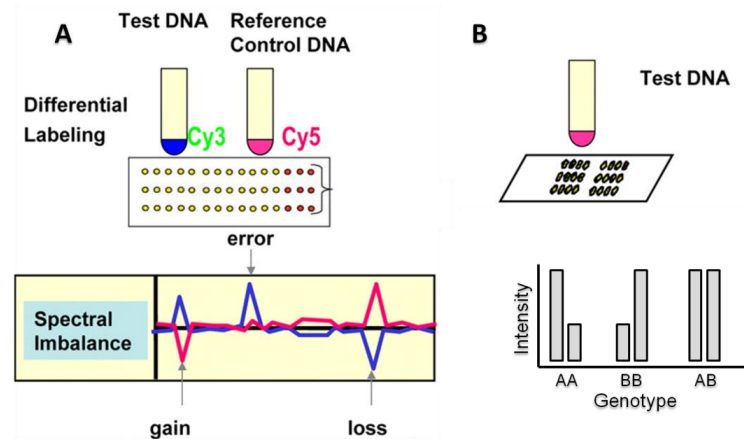
**Figure 2.13.** Genome-wide view of CNVs. Blue bars indicate reported CNVs. Red bars indicate reported inversion breakpoints. Green bars indicate segmental duplications. (Figure adapted from Database Genome Variants (hg 18)).

SNP arrays were originally designed for high-density genotyping projects with hundreds of thousands to millions of probes (e.g. in GWAS). On a SNP genotyping array, each SNP allele is represented by probes approximately 25 bp long. In contrast to the CGH array, the SNP array is oligonucleotide specific to known SNPs. They do not require reference DNA from a healthy person. A DNA sample from an individual is pre-processed and hybridized to a commercial chip. Alleles of each SNP have been defined (Figure 2.14B). Due to the high density of SNPs in the human genome, the density of oligonucleotide probes on genotyping arrays is very high, which enables the arrays to be used for CNV detection (Carter 2007). Final genotype calls and raw measurements obtained from the genotyping array have been used to construct CNV data.

### ***CNV detection using SNP arrays***

For CNV detection, signal intensities of the match and mismatch probes are compared with values from another individual (or group of individuals) and the relative copy number per

locus is determined (Yau and Holmes 2008). Highly standardized processing procedures help to reduce the variance of ratios calculated from independent hybridizations. Further noise reduction can be achieved by taking length and GC content of the probes into account. Different algorithms have been developed for the detection of CNVs in array intensity data such as PennCNV, QuantiSNP, and CNV partition (Carter 2007; Dellinger et al. 2010; Pinto et al. 2011; Valsesia et al. 2013).



**Figure 2.14.** Basic principle of array CGH and SNP array analysis. **A)** In array CGH, patient and control DNAs are labeled in different colors and co-hybridized to the array. Yellow dots on the slide indicate equal copy number for control and patient DNA. Red dots indicate the loss of DNA material in the patient. Computational analysis shows spectral imbalance which represents gain/loss of test DNA material. **B)** In a SNP array, only patient DNA is labeled and hybridized to the array. Copy number analysis is based on signal intensity; the intensity of each oligonucleotide is compared to the intensity of the same oligonucleotide in a set of standard controls.

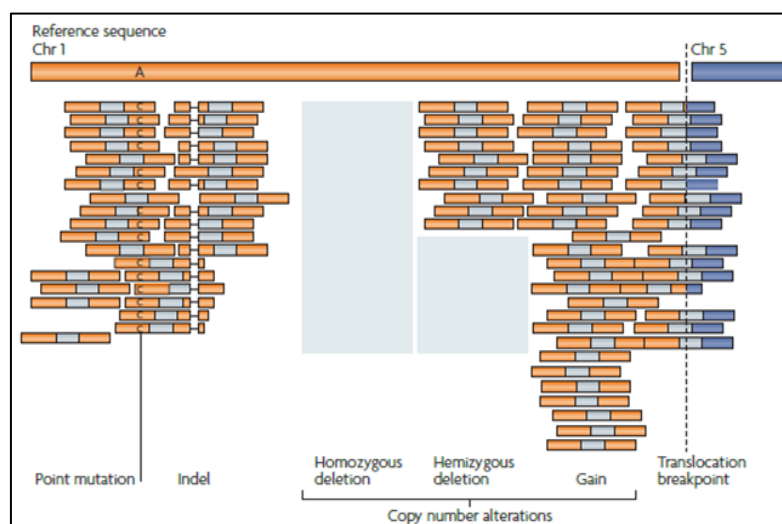
The accuracy of CNV boundaries derived from SNP arrays is influenced by multiple factors such as DNA quality, markers density, sophisticated algorithms, statistical methods, batch effects, and differences between experiments. QuantiSNP is a computational framework for detecting regions of copy number variation from BeadArray™ SNP genotyping data using an Objective Bayes Hidden-Markov Model (OB-HMM) (Colella et al. 2007). It provides probabilistic quantification of state classifications and significantly improves the accuracy of segmental aneuploidy identification and mapping, relative to analytical tools (GenomeStudio, Illumina), as demonstrated by validation of breakpoint boundaries.

The array-based approaches can only detect simple duplications or deletions, but are unable to infer more complex but copy-neutral structural variations, such as balanced inversions or translocations that can also affect genome function (Tuzun et al. 2005). These technologies show some drawbacks such as hybridization noise and limited coverage of the genome, which can lead to false interpretation of results. Thus, identified CNVs still need to be confirmed by such traditional methods as FISH, MLPA, or qPCR (see section 3.13.)

### 2.4.6. High-throughput sequencing

High-throughput sequencing, also called next generation sequencing (NGS), can be used to sequence entire genomes (whole genome sequencing; WGS) or may be constrained to specific areas of interest, including all 21,000 coding genes (whole exome sequencing; WES) or small numbers of individual genes (targeted sequencing) to describe common and rare human genetic variation and to discover novel mutations and disease causing genes. Compared to conventional Sanger sequencing, NGS allows faster and more comprehensive diagnostics in genetically heterogeneous conditions and can be more sensitive to identify low-level mosaicism. Depending on the coverage (read depth) and other quality parameters, NGS data can also be used for the detection of submicroscopic copy number changes (sequence-based approach). It is able to provide a very detailed map of genome-wide structural variants as small as 500 bp (Mills et al. 2011) (Figure 2.15). Although a rather low coverage is a current limitation to identify copy numbers (Zhao et al. 2013), further improvements of the sequencing technology and analytical tools will allow the detection of almost all known types of genetic variation by NGS in the near future.

However, since tens of thousands of genomic variants can be identified in each exome, the prioritization and interpretation of these variants are challenging. It is important to carefully consider strategies for efficiently and robustly prioritizing pathogenic variants. Many bioinformatics tools have been and are being developed to prioritize candidate disease variants from disease gene loci. The combination of prediction results with phenotypic and pedigree data as well as data from databases might be the best approach to determine the potential cause of the disease under investigation (Gilissen et al. 2012; Pabinger et al. 2013).



**Figure 2.15.** Types of genome alterations that can be detected by NGS. **Left:** point mutations (A to C), small insertions and deletions (indels) (deletions shown by a dashed line). **Middle:** changes in read depth (comparing to a normal control) is used to identify copy number changes. Grey boxes represent absent or decreased reads. **Right:** paired-ends that map to different genomic loci (chromosome 1 and 5) represent rearrangements. (Figure adapted from Meyerson et al. (2010)).

## 2.5. Validation of functionally relevant candidate genes

After promising candidate genes have been identified, the next step is to find evidence for their clinical relevance, i.e., to validate whether mutations or a certain mutation type in these genes are indeed causative. There is no single best step to success, but using multiple approaches can improve the accuracy of validation and confirm the potential causality of candidate genes. To verify plausible candidate genes, there are several approaches as described below and a few important criteria: 1) the mutation spectrum and type of mutations (missense mutations versus truncating mutations or mutations encompassing whole gene versus hotspot mutations); 2) the frequency in population-based controls; 3) in-silico prediction tools; 4) the inheritance model (biallelic mutation versus heterozygous mutation).

### 2.5.1. Recurrent findings

A gene affected by recurrent mutations is defined as a gene that harbors frequent mutations more than expected by chance. For example, recurrent mutations in the *APC* gene are common among FAP patients but rare or absent in healthy controls. For a given gene, the higher the number of mutations found in a specific patient group but not in controls, the greater is the likelihood that the gene is causative. Therefore, the simplest way to confirm causality is to look for additional mutations in large patient groups. However, germline mutations in recently identified genes causative for Mendelian conditions seem to be very rare (Meindl et al. 2010; Palles et al. 2013). Thus, mutation-negative patient groups for established familial tumor syndromes seem to be genetically very heterogeneous and, therefore, large patient cohorts are needed to validate a predisposing gene by finding recurrent mutations.

### 2.5.2. Segregation analysis

Investigating patterns of co-segregation of a mutation among affected family members is another powerful strategy to potentially increase the likelihood of a mutation being pathogenic. Co-segregation indicates causality for the phenotype or at least LD between markers and a pathogenic mutation in the respective gene (linkage marker). To be effective, reliable family histories associated with a pedigree are important. The more clearly affected family members are available, the more significant is the result. Phenotype misclassification, late onset phenotypes, mild phenotypes, and wrong paternity can lead to wrong interpretation. Segregation analysis is not effective for identifying low to moderately penetrant causative genes since in this scenario there is no clear relationship (co-segregation) between the presence of the variant in question and the development of the phenotype.



### **2.5.3. Gene expression in relevant target tissues**

Expression analysis is widely used to study whether a gene of interest is up-/downregulated. A promising candidate gene should be expressed in normal tissues which are relevant to the phenotype and the expression pattern should be consistent with the pathophysiological hypothesis of the involved genes and mutation types. For example, TSGs should be downregulated in the tumor tissue whereas oncogenes should be upregulated. The expression of candidate genes can be tested by RT-PCR, Northern blotting, Western blotting, *in situ* hybridization against mRNA in tissue sections, and expression arrays (Strachan and Read 1999).

### **2.5.4. Candidate gene approach**

The candidate gene approach, in contrast to genome-wide approaches, focuses on associations between genetic variation within pre-specified genes of interest and the phenotype. In many organisms from animals to humans, candidate gene approaches have been ubiquitously applied in gene-disease research, genetic association studies, biomarker and drug target selection (Tabor et al. 2002; Zhu and Zhao 2007).

The first critical step in conducting candidate gene studies is the choice of a suitable candidate gene that may plausibly play a relevant role in the process or disease under investigation. The selection of specific candidate genes is based on gene's functions and pathways, which should be related to the biological mechanisms of the disease (Kwon and Goate 2000). However, the main disadvantage of the selection process is that it requires information from existing well-known physiological, biochemical or functional processes. This approach is not really suited to identify novel causative genes as it is limited by the reliance on existing biological knowledge.

In the context of presumed Mendelian phenotypes, the candidate gene approach is used to look for recurrent mutations in patients to confirm the causality of candidate genes. However, in the past many candidate genes have been proposed according to functional assumptions or pathways involved, but few of them have been successfully verified and ultimately brought to endpoint usage (Zhu and Zhao 2007).

### **2.5.5. Pathway enrichment analysis/network analysis**

Enrichment analysis is a statistical method to see whether candidate genes identified in a group of patients are more frequent (overrepresented) in a certain established pathway or a functional network than expected by chance. An objective of pathway-based approaches is to connect the functional level of genes with a phenotype as the idea is that genes do not work alone but interact with other genes in functional networks. Instead of looking for a recurrent mutation in one gene, this approach aims to look for recurrent mutations in a

pathway, i.e., a set of genes (Emmert-Streib and Glazko 2011). The overrepresentation of candidate genes in a specific pathway might mean that these genes act together and are not affected in the patients just by chance.

Enrichment analysis helps to interpret data in the context of biological processes, pathways, and networks. The primary result is an enrichment score, which reflects the degree to which a gene set is overrepresented. The magnitude of the increment depends on the correlation of the genes with the phenotype. There are several web-based tools to analyze enrichment, i.e., the Database for Annotation, Visualization, and Integrated Discovery (DAVID), Gene Ontology (GO) enrichment analysis, which are used to discover enriched functionally related gene groups. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a well-known web-based tool for pathway enrichment analysis. Network analysis offers the potential for greater power of discovery and natural connections to biological mechanisms. It can confirm genes and proteins associated with the etiology of a specific disease and can help to understand mechanisms of disease (Ramanan et al. 2012).

## 2.6. Scope of the thesis

The main aim of this thesis was to identify new high penetrant causative genes in a large and well-characterized cohort of unrelated patients with clinically verified, etiologically unexplained colorectal adenomatous polyposis, a precancerous condition which is supposed to have a strong hereditary basis.

Prior to the comprehensive and laborious application of genome-wide techniques such as SNP-array based CNV analysis and targeted next generation sequencing of candidates we performed a systemic *APC* transcript analysis to uncover mutations in deep intronic regions of the *APC* gene which cannot be identified by routine diagnostics.

Afterwards, a high-resolution genome-wide CNV analysis was performed, followed by stringent filter steps to select for rare CNVs in protein coding genes which are supposed to result in truncating mutations, assuming a monogenic disease model.

Subsequently, various methods were applied to examine the causal relevance of the candidate genes including co-segregation analysis, network and pathway analysis, expression analysis, data mining, and recurrent germline point mutation and somatic point mutation analyses.

## 3. MATERIALS AND METHODS

This chapter provides detailed technical information on how this thesis was carried out. It starts with lists of databases and tools, followed by a description of available patients and how their data were analyzed.

### 3.1. Databases

- 1000 Genomes; A Deep Catalog of Human Genetic Variation:  
[www.1000genomes.org](http://www.1000genomes.org)
- APC Mutation database: [www.lovd.nl/APC](http://www.lovd.nl/APC)
- BDGP (Berkeley Drosophila Genome Project): [www.fruitfly.org/seq\\_tools/splice.html](http://www.fruitfly.org/seq_tools/splice.html)
- BLAT: <http://genome.ucsc.edu/cgi-bin/hgBlat>
- COSMIC (Catalogue of somatic mutations in cancer):  
[http://cancer.sanger.ac.uk/cancer\\_genome/projects/cosmic/](http://cancer.sanger.ac.uk/cancer_genome/projects/cosmic/)
- DAVID (Database for Annotation, Visualization and Integrated Discovery)  
Bioinformatics Resources 6.7: <http://david.abcc.ncifcrf.gov/>
- dbSNP (The Single Nucleotide Polymorphism database): [www.ncbi.nlm.nih.gov/SNP/](http://www.ncbi.nlm.nih.gov/SNP/)
- DGV (Database of Genomic Variants):  
<http://dgvbeta.tcag.ca/dgv/app/home?ref=NCBI36/hg18>
- Ensembl Genome Browser: Archive EnSEMBL release 54 - May 2009:  
<http://may2009.archive.ensembl.org/index.html>
- ESEfinder program Release 2.0: <http://rulai.cshl.edu/tools/ESE2/>
- EVS (Exome Variant Server): <http://evs.gs.washington.edu/EVS/>
- GENATLAS: <http://genatlas.medecine.univ-paris5.fr/>
- GeneCards: [www.genecards.org/](http://www.genecards.org/)
- GeneCodis (Gene annotations co-occurrence discovery):  
<http://genecodis.cnb.csic.es/>
- GWAs Catalog (A Catalog of Published Genome-Wide Associations Studies):  
[www.genome.gov/gwastudies/](http://www.genome.gov/gwastudies/)
- HGMD (Human Gene Mutation Database): <http://hgmd.org>

- KEGG (Kyoto Encyclopedia of Genes and Genomes ): [www.genome.ad.jp/kegg](http://www.genome.ad.jp/kegg)
- LOVD (Leiden Open Variation Database): [www.lovd.nl/2.0/](http://www.lovd.nl/2.0/)
- MutationTaster: [www.mutationtaster.org](http://www.mutationtaster.org)
- NCBI (National Center for Biotechnology Information): [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)
- OMIM<sup>®</sup> (Online Mendelian Inheritance in Man<sup>®</sup>): <http://omim.org/>
- PolyPhen-2: <http://genetics.bwh.harvard.edu/pph2/>
- PUBMED: <http://pubmed.com/>
- SIFT (Sorting Intolerant from Tolerant): <http://sift.jcvi.org/>
- STRING (Search tool for the retrieval of interacting genes/proteins): <http://string-db.org/newstring.cgi/>
- UCSC Genome Bioinformatics: <http://genome.ucsc.edu>
- UNIGENE: EST profile: [www.ncbi.nlm.nih.gov/unigene](http://www.ncbi.nlm.nih.gov/unigene)
- UniSTS Integrating Markers and Maps database: [www.ncbi.nlm.nih.gov/unists](http://www.ncbi.nlm.nih.gov/unists)
- VARBANK: <https://anubis.ccg.uni-koeln.de/varbank/>

### 3.2. Devices

#### Concentration measurement device

- NanoDrop<sup>®</sup> ND-1000 Spectrophotometer, Peqlab Biotechnology GmbH

#### Agarose gel electrophoresis

- UV imaging system (BioRad Gel Doc<sup>™</sup>)
- Electrophoresis horizontal gel chamber (Biorad)

#### Genotyping systems

- iScan System, Illumina<sup>®</sup> Inc.
- MassARRAY<sup>™</sup> Nanodispenser, SAMSUNG Techwin Co. Ltd. For Sequenom<sup>®</sup>
- MassARRAY<sup>™</sup> Compact Analyzer, Bruker Daltonics Inc. for Sequenom<sup>®</sup>

#### Pipetter robot systems

- Biomek<sup>®</sup> Laboratory Automation Workstations NX MC and NX S8G, Beckman Coulter GmbH

#### Sequencing devices

- Automatic 16-capillary sequencing device 3130xl Genetic Analyzer (Life Technologies, Carlsbad, CA)
- ABI 3500xl Genetic Analyzer (Life Technologies, Carlsbad, CA)

#### Thermocyclers

- MJ Research PTC-200 Thermal Cycler
- ABI Prism 7900HT Fast Real-Time PCR System (Life Technologies, Carlsbad, CA)

### 3.3. Software

- 3130xl Data Collection Software v3.0, Life Technologies
- Biomek Software 3.2, Beckman/Coulter, Fullerton, California, USA
- Cartagenia Bench Lab CNV<sup>TM</sup> software, Cartagenia, Leuven, Belgium
- Chromas v2.21, Technelysium Pty Ltd, Australia
- DesignStudio: Truseq Custom Enrichment: <http://designstudio.illumina.com/>
- Genecodis 2.0: Gene annotations co-occurrence discovery: <http://genecodis.cnb.csic.es/>
- GenomeStudio v2011.1, Illumina, San Diego, California, USA
- GenomeStudio Genotyping Analysis Module v1.9.4
- Human GenoTyping Tools: [www.mysequenom.com/Tools](http://www.mysequenom.com/Tools)
- Illumina DesignStudio: <http://designstudio.illumina.com/>
- Illumina Genome Viewer v1.9.0
- Illumina Realtime Analysis® (RTA) software
- Ingenuity pathway analysis: [www.ingenuity.com/products/ipa](http://www.ingenuity.com/products/ipa)
- Integrative Genome viewer: [www.broadinstitute.org/igv/](http://www.broadinstitute.org/igv/)
- NanoDrop ND-100 v3.3.0, Biotechnologie GmbH, Erlangen
- PLIGU (Patienten und Laborinformationssystem für genetische Untersuchungen)
- Primer3 v.0.4.0: <http://frodo.wi.mit.edu/primer3/input.htm>
- QuantiSNP v1.1 and 2.2 (Colella et al., 2007): <https://sites.google.com/site/quantisnp/>
- R software version 3.0.2: [www.r-project.org/](http://www.r-project.org/)
- Sequence Detection Software; SDS 2.2.2, Life Technologies
- SeqPilot v4.0.1, JSI medical systems GmbH, Germany

- Sequencing Analysis Software v5.2.0, Life Technologies Corporation, Carlsbad, California, USA
- Sequenom's MassARRAY Designer software
- SNP Annotation and Proxy Search (SNAP): [www.broadinstitute.org/mpg/snap/](http://www.broadinstitute.org/mpg/snap/)
- Typer v3.4 and v4.0, Sequenom

### 3.4. Commercial reagents

- 100 bp DNA ladder (BioEngland)
- 5 x Big Dye® Terminator v1.1 sequencing buffer (Life Technologies)
- Agencourt AMPure® XP PCR purification kit (Beckman Coulter)
- Agencourt CleanSEQ® dye-terminator removal kit (Beckman Coulter)
- Amplitaq Gold® PCR Master Mix (Life Technologies)
- Big Dye® ready reaction mix v3.1 (Life Technologies)
- DyeEx 2.0 Spin kit (Qiagen)
- Expand 20 kbPlus PCR system (Roche Diagnostics GmbH, Mannheim, Germany)
- First strand cDNA of human adult colon mucosa (Amsbio)
- HighPure PCR Product Purification Kit (Roche Diagnostics)
- Human MTC™ Panel I & II (Clontech)
- Infinium II Whole-Genome Genotyping Kit (Illumina Inc.)
- PAX gene blood RNA kit (Qiagen)
- Power SYBR Green® PCR Master Mix (Life Technologies)
- QIAamp DNA FFPE tissue kit (Qiagen)
- QIA quick PCR purification kit (Qiagen)
- REDTaq ReadyMix PCR Reaction Mix with MgCl<sub>2</sub> (Sigma Alrich)
- SEQUENOM® iplex® Gold Chip and Reagent Kit (Sequenom)
- Super Script First-Strand Synthesis System for qRT-PCR (Invitrogen)
- TaqMan gene expression assays: Human *CTNNB1* and Human *MUTYH* (Life Technologies)
- TaqMan endogenous controls: Human Cyclophilin (hu-CYC) and Human Beta-2-microglobulin (hu-β2M)( Life Technologies)

- TaqMan® gene expression master mix (2X) (Life Technologies)
- TruSeq® Custom Enrichment kit (Illumina Inc.)
- TruSeq® DNA HT Sample Preparation Kit (Illumina Inc.)

### 3.5. Study samples

#### 3.5.1. Initial patient cohort

Patients or blood/DNA samples were primarily referred from all parts of Germany to the Institute of Human Genetics, Bonn, or the Medical Genetics Center (MGZ), Munich, Germany, because typical or attenuated adenomatous polyposis was suspected. All patients were screened during routine diagnostics for germline mutations in the *APC* and *MUTYH* genes by complete Sanger sequencing of the coding regions and the flanking exon-intron boundaries (~30-40 bp) as described (Aretz et al. 2006). MLPA analysis (MRC-Holland, Amsterdam, Netherlands) was performed to screen for large genomic deletions or duplications of the *APC* gene.

**Table 3.1.** Clinicopathological and genetic features of the patients included

Study project	CNV analysis	Targeted NGS		
Phenotype	FAP	FAP	HNPCC	Total
<b>No. of patients</b>	229	145	47	192
<b>Gender (male/female)</b>	132/97	83/62	22/25	105/87
<b>Mean age at diagnosis (years)</b>	45	46	52	47
<b>Range (years)</b>	12-78	12-78	30-86	47-82
<b>No. of colorectal adenomas</b>				
< 100 adenomas	142 (62%)	99 (68%)		
> 100 adenomas	42 (18%)	20 (14%)		
Multiple/numerous polyps	42 (18%)	24 (17%)		
unknown	3 (1%)	2 (1%)		
<b>Colorectal cancer</b>	77 (34%)	55 (38%)	41 (87%)	96 (49%)
<b>Extracolonic lesions *</b>				
yes	34(15%)	18 (12%)		
no	195 (85%)	127 (88%)		
<b>Family history</b>				
Familial	34 (15%)	19 (13%)	47 (100%)	66 (34%)
Sporadic	188 (82%)	123 (85%)		123 (64%)
Unclear/unknown	7 (3%)	3 (2%)		3 (2%)



All index patients with no detectable germline mutation in the *APC* or *MUTYH* genes were contacted by clinical geneticists of the Institute of Human Genetics, Bonn, either directly or via their responsible clinician and asked to participate in the present study. Clinical information and family history of patients were obtained during genetic counseling sessions, from a questionnaire, through telephone interviews and from medical records. Patients were asked to enrol in the study and to give their informed consent. Missing medical notes and histopathology records were requested from general practitioners, medical specialists, hospitals, and institutes. Affected relatives were informed about the study by the patients and were asked to participate in the study. The study was approved by the ethics review board of the Faculty of Medicine, University of Bonn, Germany.

### 3.5.2. NGS validation cohort

192 patients (87 females, 105 males) were selected for the study. One hundred of them came from the CNVs study cohort. Forty-five adenomatous polyposis patients and an additional 47 subjects with suspected Lynch syndrome (Table 3.1) were subsequently included. In this latter group, the Amsterdam I or II criteria for Lynch syndrome (Vasen et al. 1991; Vasen et al. 1999) were met; the tumors showed microsatellite stability and normal immunohistochemistry. Moreover, no mutations in *APC*, *MUTYH*, *EPCAM*, and mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*) could be identified by extensive mutation screening including Sanger sequencing, MLPA, and transcript analysis. The recruitment process and inclusion criteria of these patients into the study were the same as with the initial patient cohort.

### 3.5.3. Heinz Nixdorf RECALL (HNR) study controls

The Heinz Nixdorf RECALL (Risk Factors, Evaluation of Coronary Calcium and Lifestyle) cohort study (HNR controls) is an ongoing longitudinal study taking place at the University Hospital of Essen, and analyzes the population-based incidences of cardiovascular diseases ([www.recall-studie.uni-essen.de/](http://www.recall-studie.uni-essen.de/)). The entire sample comprises about 4,800 men and women aged from 45 years to older.

Samples used in this study are of German descent and were chosen according to their genotyping chip. Using the same platform can exclude sampling biases and the use of platform-specific references makes it possible to account for platform-specific artifacts, thus this study included only samples which were hybridized on the Omni-1Quad SNP-array. According to the inclusion criterion that the standard deviation of the logR ratio must be lower than 0.3, 531 HNR controls (255 females, 276 males) were included in the CNV study.

### 3.5.4. GWAS replication study

To replicate the results of a GWAS in unexplained adenomatous polyposis, recently performed by the research group (results not shown here, and not yet published), the most promising SNPs were genotyped in a large group of Dutch patients and controls, provided by our collaborators at Leiden University, The Netherlands. DNA samples from 950 subjects comprising 379 adenomatous polyposis patients with no *APC* or *MUTYH* germline mutation, and 570 anonymous controls were analyzed. Age at diagnosis of patients vary from 4-82 years (median 52) while ages of the control cohort varied from less than ten up to 88 years (median 44). The basic characteristics and colorectal phenotype of the patient cohort is summarized in table 3.2.

**Table 3.2.** Clinicopathological and genetic features of the patients

Study cohort	Patient
Number of sample	380
Gender (male/female)	219/160 (unknown 1)
Median age at diagnosis (years)	52
Range (years)	4-82
<b>Pathological report</b>	
polyps < 10	68 (18%)
polyps 10-50	194 (51%)
polyps 50-100	58 (15%)
polyps > 100	35 (9%)
Not available	25 (7%)
<b>Family history</b>	
Familial	204 (54%)
Sporadic	176 (46%)

## 3.6. DNA and RNA preparations

### 3.6.1. DNA extraction using desalting method

Using a standard salting-out procedure, genomic DNA was extracted from peripheral EDTA-anticoagulated blood samples (Miller et al. 1988). Ten ml of EDTA treated blood was transferred into a 50 ml Falcon tube and cell lysis buffer (saturated NaCl [6M]) was added up to 45 ml and mixed well. The tube was incubated on ice for 15 minutes before being centrifuged for 15 minutes at 1500 rpm and 4°C. Supernatant was discarded and pellet was air-dried for 5-10 minutes. Five ml of nuclear lysis buffer (10mM Tris-HCl pH 8.2, 400 mM NaCl, 2mM Na<sub>2</sub>EDTA, pH 8.2) was added into the tube, followed by 250 µl of protein lysis buffer (10 mg/ml proteinase K). The solutions were mixed well, then 250 µl of 10% SDS was added and incubated overnight at 37°C to complete the lysis reaction. Two ml of saturated NaCl was added into the sample tube and centrifuged at 4000 rpm for 10 minutes at room

temperature. Supernatant was transferred to a new 50 ml Falcon tube and mixed with 5 ml isopropanol. A visible DNA strand was removed with a spatula and transferred to a 1.5 ml microcentrifuge tube containing 70%EtOH to dehydrate the DNA strand. The dried DNA was transferred to a new 1.5 ml microcentrifuge tube containing 200-500 µl of TE-4 buffer. The DNA was incubated at room temperature at least 2 hours to dissolve the DNA completely before quantitation. It can then be stored at 4°C for months.

### **3.6.2. Formalin fixed paraffin embedded (FFPE) tissue DNA isolation**

Adenomas were obtained during an operation and were fixed and preceded follow standard formalin-fixation and paraffin-embedding procedures and the tissue blocks were kept at room temperature.

The formalin fixed paraffin embedded (FFPE) tissues were cut with a thickness of 10 micron. One section of each block was stained with Hematoxylin and Eosin (H&E) to indicate the tumor location. Up to 8 sections with a surface area of tumor up to 250 mm<sup>2</sup> were combined in one preparation. DNA was extracted with the QIAamp DNA FFPE tissue kit (Qiagen) followed the manufacturer's protocol. Briefly, paraffin was dissolved in xylene, then sample was lysed with proteinase K and incubated at 90 °C to reverse formalin crosslinking. DNA was binded to a membrane provided in the kit while residual contaminants were washed away. Then the DNA was eluted from the membrane and kept at 4°C.

### **3.6.3. RNA extraction using the PAX gene kit**

RNA was extracted from approximately 2.5 ml of venous blood collected in a PAXgene Blood RNA Tube (Becton Dickinson) containing RNA stabilization reagent. In principle, the PAX gene blood RNA kit (Qiagen) uses a membrane-based isolation and purification system in the form of a 'spin column'. The PAXgene Shredder spin column is used to homogenize the cell lysate and remove residual cell debris. The PAXgene RNA spin column with silica membrane will bind to RNA whereas contaminants pass through the membrane. All reagents for RNA reactions were provided in the kit. The RNA extraction protocol is described in the manufacturer's handbook. Briefly, before starting the RNA extraction process, all equipment and the working area must be cleaned with 99% chloroform to reduce contamination with RNases. Whole blood was collected in the PAX gene blood RNA tube and incubated at room temperature overnight to complete the blood cells lysis. The pellet was washed and resuspended. The resuspended pellet was incubated in optimized buffers together with proteinase K to bring about protein digestion. After an additional centrifugation through the PAXgene Shredder spin column, the supernatant of the flow-through fraction was transferred to a fresh microcentrifuge tube. Ethanol was added to adjust binding conditions, and the lysate was applied to a PAXgene RNA spin column. Remaining contaminants were removed in several efficient wash steps. Between the first and second wash steps, the membrane was

treated with DNase I to remove trace amounts of bound DNA. After the wash steps, RNA was eluted in elution buffer and heat-denatured at 65°C. The reaction was stopped by immediately placing it on ice. The RNA was quantified by UV absorbance (260/280 nm) and stored at -70°C.

#### **3.6.4. Determination of concentration and quality**

A spectrophotometer is commonly used to determine the concentration of nucleic acid based on the Beer-Lambert equation. The Beer-Lambert law provides a relationship between the amount of the light absorbed and the concentration of the absorbing molecule. The absorbance is used to convert optical density (OD) to concentration. The value of 1 unit of OD is equivalent to 50ng/μl for DNA and 40 ng/μl for RNA. The DNA concentrations and purity are commonly determined by measuring the ratio of the UV absorbance at 260 nm and 280 nm. Based on the fact that OD for DNA at 260 nm is twice that at 280 nm, the cleaned DNA has an OD-260/OD-280 ratio of between 1.8 and 2.0, and around 2.1 for an RNA sample.

For this thesis, the measurement was performed with a Thermo Scientific NanoDrop™ ND-1000 Spectrophotometer, which requires only 1 μl of DNA for the measurement.

#### **3.6.5. First-strand cDNA synthesis**

cDNA represents a more convenient way to work with the coding sequence because RNA is very unstable, fragile, and easily degraded. First strand cDNA synthesis, or reverse transcription (RT), is a process, which transcribes single-stranded RNA into complementary DNA (cDNA). It is facilitated by a reverse transcriptase enzyme, RNA-dependent DNA polymerase, which is typically of Avian Myeloblastosis Viral (AMV) origin. This step is a required procedure prior to amplification by DNA polymerases.

There are three different types of primers: 1) oligo-dT primer is used when the mRNAs have a poly-A tail; it anneals to all mRNA simultaneously. 2) Sequence-specific primer can be used to produce specific cDNA from a particular mRNA. 3) Random primer is able to produce pieces of cDNA scattered all over the mRNA. For this thesis, random primers were used to generate first-strand cDNA.

One μg of RNA from lymphocytes was reverse transcribed into a first cDNA strand by random hexamer-primed reverse transcription. This was achieved with the Super Script First Strand Synthesis System for qRT-PCR (Invitrogen) according to the protocol provided by the manufacturer as outlined below.

One μg of RNA was diluted in 8 μl of DEPC-treated water. One μl of 10 mM dNTP mix and one μl of Random hexamers (50 ng/μl) were added to the tube, followed by brief centrifugation and incubation at 65°C for 5 minutes, then immediately placed on ice for 2

minutes. In a separate tube, 2X reaction mix was prepared. Volume and concentration are shown in table 3.3.

The 2X RT reaction mix was added to the RNA tube, which was then incubated at room temperature or 25°C for 2 minutes. One µl of SuperScript™ II RT was added into each tube and tubes were incubated at room temperature for another 10 minutes, temperature was increased to 42°C for 50 minutes, and the reaction was terminated after 15 minutes at 70°C temperature. After termination, the reaction material was immediately cooled down on ice, and then briefly centrifuged. One µl of RNase H was added to remove remaining template RNA. The reaction mix was incubated at 37°C for 20 minutes. The first-strand cDNA was either immediately used for qPCR or stored at -20°C.

**Table 3.3.** Solutions and concentrations for reaction mix preparation

<b>Solution</b>	<b>Volume</b>
10X RT buffer	2 µl
25 mM MgCl <sub>2</sub>	4 µl
0.1 M DTT	2 µl
RNaseOUT™ (40 U/µl)	1 µl
Total volume	9 µl

### 3.7. Polymerase Chain Reaction (PCR)

#### 3.7.1. Basic principle

The Polymerase Chain Reaction (PCR), invented by Kary Mullis in 1983, is a technique for amplifying specific target DNA. The basic principle is the cyclic change of different temperatures to promote an enzymatic amplification of specific DNA. To start the reaction the temperature is raised to 95°C to melt double-stranded DNA into single strands. The temperature is then lowered to 50-60°C to allow primers to bind to target DNA. Thus the polymerase enzyme has somewhere to bind and can begin synthesizing a complementary sequence of the DNA strands with deoxynucleotide triphosphates (dNTPs). The optimal temperature for the polymerase is 72°C, which allows the enzyme to work fast. The melting of double strand DNA is called “Denaturation step”, while the binding of primer is called “Annealing step”, and the synthesizing of the complementary sequence is called “Extension step”. These three steps are repeated 30 to 40 times and the amount of DNA is increased exponentially. At the end of the amplification, the product can be detected on an agarose gel.

#### 3.7.2. Primer design

In this thesis, Primer3 (V.0.4.0), an online program tool (<http://frodo.wi.mit.edu/primer3/input.htm>), was used to design primers for quantitative real-time PCR (qPCR), Sanger

sequencing, and expression analysis. DNA sequences were called by the *Ensembl* Genome Browser; release 54 – May 2009, based on the NCBI reference sequence [RefSeq] build 36. The length of primers was set at 18-27 bases, melting temperature was 57-61°C, and GC content was 20-60%. When the software was unable to design proper primers, they were designed manually following these criteria:

- Length: 18-22 bp
- Melting temperature: 58-60°C,  $\Delta T_m$  should not be higher than 2°C
- GC content: 40-60%
- GC clamp: put GC at 3' end if possible, avoid > 2 GC sequences in the last 5 bases at the 3' end
- avoid a repeat of > 4 di-nucleotides
- avoid polymorphism

Amplicon length for PCR was set at 400-600 bp while the length for the quantitative PCR product was set at 120-150 bp and the length for Sanger sequencing was at most 500 bp. Primers and amplicons were blasted on the UCSC Genome Browser to verify the specifications. The primers were then synthesized at Metabion (Martinsried, Germany) in standard quality.

### 3.7.3. PCR reaction components

Reagent	Volume (µl)	Final concentration
10x PCR buffer with MgCl <sub>2</sub>	2.5	1X
10x dNTP mix (10 mM each)	0.5	0.4 µM each
Taq polymerase [5 U/µl]	0.2	1 U/reaction
DNA [10 ng/µl]	2	20 ng/reaction
10 µM For primer	1	0.4 µM
10 µM Rev primer	1	0.4 µM
H <sub>2</sub> O	17.8	
Total	25	

### 3.7.4. Cycling step

Reaction	Temperature	Duration	Cycle
Initial denaturation	95°C	5 min	1 X
Denaturation	95°C	30 sec	35 X
Annealing	56°C	30 sec	
Extension	72°C	1 min	
Final extension	72°C	10 min	1 X
Storage	4°C	∞	1 X

### 3.7.5. Agarose gel electrophoresis

Agarose is a polysaccharide which, after solidifying, forms a three-dimensional network that allows for migration of DNA molecules in a buffer. Briefly, if an anode is attached to the system, negatively charged DNA migrates through the agarose pores based on the effect of molecular sieving. The DNAs can then be separated according to their sizes as short molecules move faster and thereby further than long molecules.

2-20 µl of PCR product was loaded on 2% agarose gel. A 100 bp-DNA ladder (BioEngland) was used as a molecular weight marker. The electrophoresis system was set at 120 volt and run for 1.30 hours to separate the PCR product. The result of a gel electrophoresis run was analyzed after ethidium bromide staining using a UV imaging system (BioRad Gel Doc™) and visualized with (BIORAD).

### 3.7.6. PCR product purification

**QIA Quick PCR Purification:** After specificity and quantity of the products were checked by running agarose gel electrophoresis, the amplified PCR product was purified using the QIA quick PCR purification kit (Qiagen). This purification is based on the silica-membrane spin column. All purification steps, that is, DNA binding to the filter membrane, washing steps, and the DNA elution were performed step by step following the company manual. The cleaned product was used for validation of the mutation by Sanger sequencing.

**AmPure purification:** Agencourt AMPure XP PCR purification kit was used to purify PCR product. It is based on Solid Phase Reversible Immobilization (SPRI®) with magnetic bead-based technology. DNAs bind to the magnetic beads and the beads are adhered to the walls by a magnetic plate while other molecules such as primer dimers, dNTPs, and so on are washed away. After washing steps, the DNA is eluted and detached from the beads.

Before using, the AmPure bottle was gently shaken to resuspend the magnetic particles, then 36 µl of AmPure buffer containing the magnetic beads in solution was added into the PCR reaction and well mixed by pipetting up and down 10 times. The PCR reaction plate

was placed on the magnetic plate for 5-10 minutes to separate beads from solution and supernatant was discarded. Two-hundred  $\mu$ l of 70% EtOH was added into the reaction and incubated at room temperature for 30 seconds, then the supernatant was discarded and this washing step was repeated for a total of 2 washes. The purified PCR product was air-dried for 20-30 minutes to completely remove ethanol and dissolved in 40  $\mu$ l of elution buffer (TE<sup>-</sup>4).

**Purification PCR product from agarose gel:** The specific band PCR product on the gel was removed from the gel and purified with the High Pure PCR Product Purification Kit (Roche Diagnostics). All processes were performed following manufacturer's protocol as described. To excise the specific band, we used only a new blade. Binding buffer was added into a fresh 1.5 ml microcentrifuge tube which contained the excised specific agarose gel, and incubated at 56°C for 15 minutes to melt the gel, followed by adding isopropanol. Then the product was transferred into a filter tube and was centrifuged at 13000 rpm for 1 minute, then supernatant was discarded. The rest of the product mix was transferred into the filter tube and the centrifuge step was repeated. After discarding the supernatant, the filter tube was filled with washing buffer, centrifuged, and supernatant discarded twice before contents were transferred to a new 1.5 ml microcentrifuge tube. Sixty  $\mu$ l of elution buffer was added into the filter tube and the filter tube was incubated at room temperature around 10 minutes to complete the elution before being centrifuged at 13,000 rpm for 1 minute. Purified DNA in the 1.5 ml microcentrifuge tube was stored at 4°C and the filter tube was discarded.

### 3.8. Sanger sequencing

#### 3.8.1. Basic principle

Sanger sequencing, developed by Fred Sanger et al. in the mid 1970's, is a tool to detect the alteration of nucleotides of a DNA sequence (Sanger and Coulson 1975). It is the most reliable method for detecting sequence variations. The sequencing reactions are analogous to the PCR reactions for replicating DNA. The template DNA pieces are replicated, incorporating normal nucleotides, but occasionally and at random di-deoxynucleotides (ddNTPs) are taken up. The Sanger technique uses ddNTPs, which are essentially the same as nucleotides except they contain a hydrogen group instead of a hydroxyl (-OH) group at 3' carbon. Because they lack a 3' OH, nothing can be added to the chain once a ddNTP has been added; therefore, the replication is stopped. Sooner or later all of the copies will get terminated by the ddNTPs, but each time the enzyme makes a new strand; the place where it was stopped will be random. Because of their different lengths, running at different rates during electrophoresis, their order can be determined.

Each type of ddNTP emits colored light of a characteristic wavelength. The color band is recorded on a simulated image and interpreted by a computer program, which automatically



prints out a chromatogram as well as the sequence. The colors represent the four bases: blue is C, black is G, red is T, and green is A.

### 3.8.2. Reaction components

Reagent	Volume (μl)
5x Big Dye Terminator v1.1 sequencing buffer	3.75
Big Dye - ready reaction mix v3.1	0.5
5 μM Primer For/Rev	1
Purified PCR product	1
H <sub>2</sub> O	13.75
Total	20

### 3.8.3. Cycling step

Reaction	Temperature (°C)	Duration	Number of cycle
Initial denaturation	96	1 min	1 X
Denaturation	96	10 sec	25 X
Annealing	50	5 sec	
Extension	60	4 min	
Storage	12	∞	1 X

### 3.8.4. Cycle sequencing product cleaning

**DyeEx 2.0 Spin Kit (Qiagen):** The DyeEx 2.0 Spin kit (Qiagen) with prehydrated DyeEx gel-filtration material was used to remove unbounded ddNTPs from the sequencing reaction before loading the tagged products onto a capillary sequencer. When sequencing reaction mixtures are applied to DyeEx columns, dye terminators diffuse into the pores and are retained in the gel-filtration material, while labeled DNA fragments are excluded and recovered in the flow-through.

A quick centrifugation step was applied to remove storage buffer from the column, the sequencing samples were loaded, and a second centrifugation step was performed at 3000 rpm for 3 minutes to remove unincorporated dye terminators. Samples were then ready for loading onto a capillary sequencer.

**CleanSEQ purification:** CleanSEQ® is the Agencourt dye-terminator removal kit. As is the AMPure kit, it is based on SPRI® paramagnetic bead technology, where the nucleic acids are immobilized onto paramagnetic micro particles using specific buffer conditions.

Ten  $\mu\text{l}$  of CleanSEQ buffer, containing magnetic beads, was added into each reaction, followed by 62  $\mu\text{l}$  of 85% EtOH, and then mixed by pipetting up and down 7 times. The PCR reaction plate was placed on the magnetic plate for 3 minutes to separate beads from solution, and then supernatant was discarded. One-hundred  $\mu\text{l}$  of 85% EtOH was added into the reaction and incubated at room temperature for 30 seconds and then the supernatant was discarded. These washing steps are to remove unincorporated dyes, nucleotides, salts and contaminants. The purified product was air-dried for 10-20 minutes to completely remove ethanol before adding 40  $\mu\text{l}$  of  $\text{H}_2\text{O}$  into the reaction.

### **3.8.5. Capillary electrophoresis**

Capillary sequencing is an accurate nucleic acid analysis. DNA passes through the detection cell and a laser beam simultaneously illuminates the capillaries from both sides of the array. To accomplish this, light from a single laser source is split using optical elements to form a dual pathway. The emitted fluorescent light is collected, separated by wavelength, and focused onto a charge-coupled device (CCD). When the fluorescent light has been collected and dispersed across the CCD, the data are transferred to the instrument computer where they are transformed by chemometric algorithmic processing into 4-dye electropherograms.

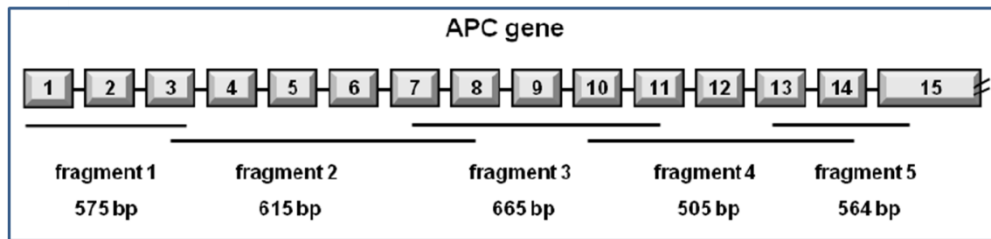
This thesis used an automated 16-capillary sequencing device, 3130xl Genetic Analyzer (Life Technologies). One  $\mu\text{l}$  of purified product was diluted in 9  $\mu\text{l}$   $\text{H}_2\text{O}$  before it was put into the sequencer.

## **3.9. APC transcript analysis**

To uncover aberrant transcripts, which might be caused by intronic mutations outside the routinely screened regions, we performed a systematic mRNA analysis of the *APC* gene in 125 apparently unrelated mutation negative patients with clinically verified adenomatous polyposis and negative *APC* and *MUTYH* mutation.

### **3.9.1. Primer design**

Primers of coding regions of the *APC* gene were designed with the primer 3 online tool (see section 3.7.2). The cDNAs were fragmented into 5 overlapping fragments spanning exons 1-15A (c.-138 to c.2625) (Figure 3.1).



**Figure 3.1.** Schematic overview of the *APC* gene (coding exons 1-15A) and the five overlapping fragments used for cDNA analysis

### 3.9.2. cDNA analysis

Specific fragments of cDNA from 125 included patients were amplified using 5 primer pairs (Table A1). The PCR reactions and cycles are described in section 3.7. Twenty µl of PCR products were loaded on agarose gel to check sizes of amplified products in comparison with the commercial marker, 100 bp-DNA ladder (BioEngland), and to compare the pattern of the bands with those in 10 anonymous controls.

If the patient's agarose gel illustrated the pattern and the size of the product differently from those in the control cohort, the specific different bands on the gel were cut out of the gel and purified (see section 3.7.6.) for re-amplification with the same primer pair to increase the yield of the product. The re-amplified product was cleaned with the QIA Quick PCR Purification Kit (Qiagen) (see section 3.7.6.) before being tagged with fluorescent dyes for sequencing.

### 3.9.3. Sanger sequencing

Both forward and reverse primers were used for cycle sequencing. The capillary electrophoresis was performed on an ABI 3500xl Genetic Analyzer (Life Technologies). For details, see section 3.8.

### 3.9.4. Data analysis

The sequencing results were visualized using Chromas software version 2.21 (Technelysium Pty Ltd, Australia) and SeqPilot software version 4.0.1 (JSI medical systems GmbH, Germany). The cDNA bases were numbered according to the *APC* reference sequence in GenBank NM\_000038.5, where +1 corresponds to the A of the ATG translation initiation codon. All abnormal results were confirmed at the genomic level in an independent experiment.

### 3.9.5. Genomic DNA analysis

If an exonised intronic insertion or exonic deletion was detected, the corresponding region was sequenced at the genomic DNA level with primers flanking that particular region (Table A2).

### 3.9.6. Haplotype analysis

A haplotype refers to a set of alleles at genetic markers (microsatellites, SNPs) that are inherited together on a chromosome. The haplotype “phase” refers to the determination of a haplotype or the placement of alleles together along a chromosome. Affected family members are likely to share the same haplotype.

In this thesis, haplotype analysis was performed with a panel of seven microsatellite markers flanking the *APC* region. The order of markers on chromosome 5q is: CEN – D5S134 – D5S492 – D5S1965 – *APC* – D5S346 – D5S656 – D5S2001 – D5S421 – TEL. They span around 2.8 Mb. The primer sequences are given in table A3.

For each microsatellite analysis, primers were picked from the UniSTS Integrating Markers and Maps database ([www.ncbi.nlm.nih.gov/unists](http://www.ncbi.nlm.nih.gov/unists)). Five ng of genomic DNA was amplified with forward primer labelled with FAM fluorescent dye and with the specific unlabeled reverse primers and Amplitaq Gold® PCR Master Mix using standard protocol. The annealing temperature was adjusted following the melting temperature of primers. The PCR amplicons were analyzed and separated by size with capillary electrophoresis. The expected size of the products is given in table A3.

### 3.9.7. In-silico analysis

Splicing efficiencies of the normal and mutant sequences were calculated using the splice prediction program NNSPLICE 0.9 from BDGP (the Berkeley Drosophila Genome Project; [www.fruitfly.org/seq\\_tools/splice.html](http://www.fruitfly.org/seq_tools/splice.html)) and with the method of Shapiro and Senapathy (1987). The influence of base substitutions on a putative Exonic Splicing Enhancer (ESE) site was checked with the ESEfinder program (Cartegni et al. 2003). The location of inserted sequences at the genomic level was determined by BLAST analysis.

## 3.10. Genome-wide SNP array hybridization

### 3.10.1. Genotyping based on BeadArray Technology (Illumina®)

The basic principle of DNA microarrays is a binding or hybridization assay. Applications of the arrays are diverse, such as SNP detection and scoring, gene expression and mutation analysis of large genes. They have been used extensively for linkage disequilibrium studies

for investigating genetic associations with disease, for genome wide association studies (GWAS), and for copy number variation (CNV) studies.

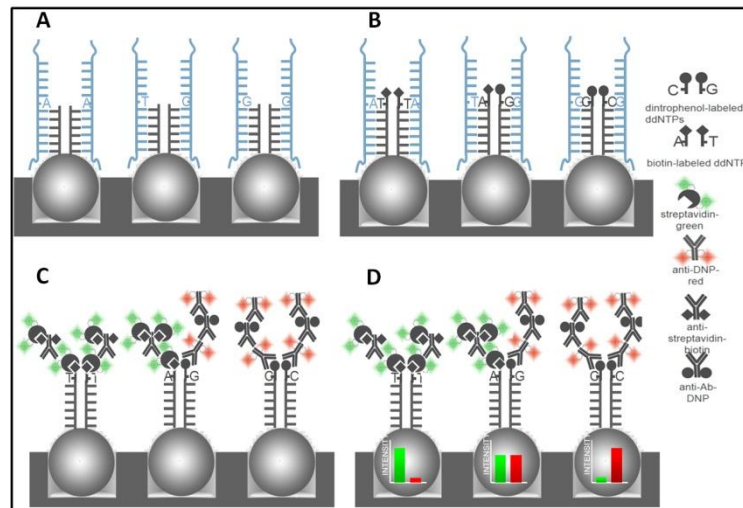
The Illumina BeadArray technology is based on 3 micron silica beads in microwells (Fearon 2011). Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in the reaction. The target is amplified, hybridized on the array, and fluorescently labeled. The array is read by exciting the fluorescence signal with a laser scanning each spot or imaging the entire array (Figure 3.2).

For this thesis, the Human Omni 1-Quad Bead chip (Illumina) was used to identify copy number changes in all samples. The chip contains 1,140,419 markers including both SNP and CNV probes; 618,959 of them are SNP probes with 10 kb of RefSeq genes and 32,110 markers are nonsynonymous SNPs (NCBI annotated). The average genomic distance between adjacent markers is 2.4 kb. The log R deviation is below 0.30.

### 3.10.2. Protocol

Amplification, hybridization, and extension steps were performed following the protocols of Infinium® HD Assay Super, recommended by the manufacturer (Illumina) as described below.

Briefly, each whole genome amplification reaction requires 200 ng of DNA (4 µl of 50 ng/µl DNA) and enables up to 1000-fold amplification. The amplified DNA was enzymatically fragmented using end-point fragmentation and was then precipitated with isopropanol at 4°C. The precipitated DNA was resuspended in hybridization buffer and then hybridized onto a BeadChip. The loaded BeadChip was incubated overnight at 48°C. The amplified and fragmented DNA samples anneal to locus-specific 50-mers (covalently linked to one of up to one million bead types) during hybridization. Unhybridized and non-specifically hybridized DNA was washed away. The BeadChip was labeled with ddNTPs; dinitrophenol (DNP)-labeled ddNTPs (C/G), and biotin-labeled ddNTPs (A/T), to extend the primer hybridized to the DNA. This step is known as 'single base extension'. The haptens are stained with Streptavidin or anti-DNP immunoglobulin. Signal amplification is used in combination with anti-streptavidin or antibody to the anti-DNP immunoglobulins conjugated to a fluorescent reporter (Figure 3.2). The iScan Reader (Illumina) was used to scan the BeadChip. The reader uses a laser to excite the fluorophore of the single-base extension product on the beads of the beadchip. Light emissions from the fluorophore are recorded in high-resolution images. Data from these images were analyzed with the GenomeStudio Genotype Module (Illumina) and export to a file containing the SNP locus, genotypes and quantified fluorescent signal intensities (Xraw, Yraw).



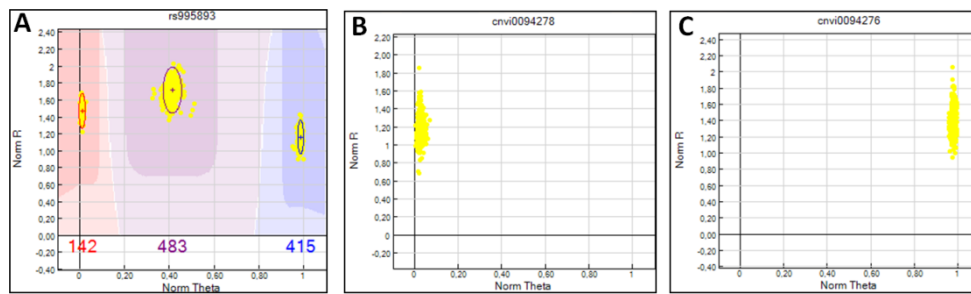
**Figure 3.2.** Illustration of Single Base Extension and staining steps: after hybridization (A), ddNTPs are labeled (B), and stained with fluorescent dyes (C). The fluorescence signals are recorded and the intensity is analyzed by a software tool (D). (Pictures adapted from [http://barleyworld.org/sites/default/files/illumina\\_snp\\_genotyping.pdf](http://barleyworld.org/sites/default/files/illumina_snp_genotyping.pdf))

### 3.10.3. Bead decoding

Decoding is achieved by a series of sequential hybridizations. Oligonucleotides are usually 80 bases long. The decoding process (Gunderson et al. 2004) uses the first 50 bases of the oligonucleotide for detection of the beads while the remaining 30 bases are complementary to the genomic target sequence. The 3' end of the complementary sequence of the oligonucleotide ends at a position before the analyzed SNPs. Each bead type is defined by a unique DNA sequence, which is recognized by a complementary decoder. The fluorescence signal is read by imaging the entire array. If one assigns a number to each state – 0 to blank, 1 to green, and 2 to red, then each cycle of the process generates one of three digits. The array is then dehybridized and the rehybridization process is repeated until there is sufficient data to unambiguously determine the identity of each bead.

### 3.10.4. Quality control of raw data

The GenomeStudio Genotyping Module v2011.1 (Illumina Inc) was used for the analysis of the genotyping assay collected by the iScan system (Figure 3.3). This module enables efficient genotyping data normalization, genotype calling, clustering, data intensity analysis, loss of heterozygosity (LOH) calculation, and copy number variation (CNV) analysis. The software calculated a SNP call rate, which represents the percentage of the number of markers detected by the scanner. For efficient CNV identification, we set a call rate threshold of  $\geq 97\%$ .



**Figure 3.3.** Genoplots from GenomeStudio Genotyping Analysis Module with data of a SNP marker (A) with 3 different genotypes (red = AA, purple = AB, blue = BB) and data of CNV markers (B and C) showing only 1 genotype each. Genotypes are called for each sample (dot) by their signal intensity (norm R) and allele frequency (Norm Theta) relative to canonical cluster positions (dark shading) for a given SNP marker.

### 3.11. Identification of putative CNVs

#### 3.11.1. Final reports

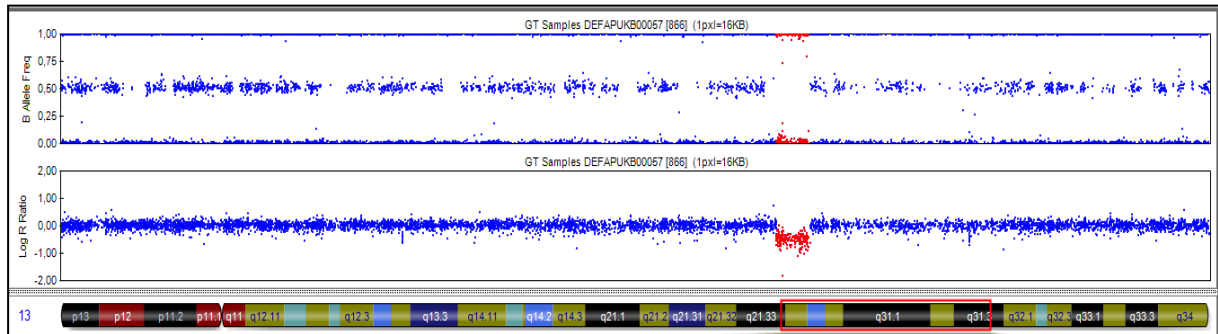
Genome Studio Files were divided into two sets according to the dates at which blood samples were received. The first Genome Studio file contains data of 181 patients and 531 controls (total  $n = 732$ ). The second file contains data of 229 patients and 531 controls, including the samples of the first file (total  $n = 760$ ).

To call putative CNVs, two final reports were created by the Illumina GenomeStudio Genotyping module v2011.1 (Illumina Inc.). The first final report was created from the first Genome Studio file for 181 patients and 531 controls and the second report was created from the second Genome Studio file for the additional 48 patients. The final report indicates position, log R ratio (LRR), and B Allele Frequency (BAF) of all markers.

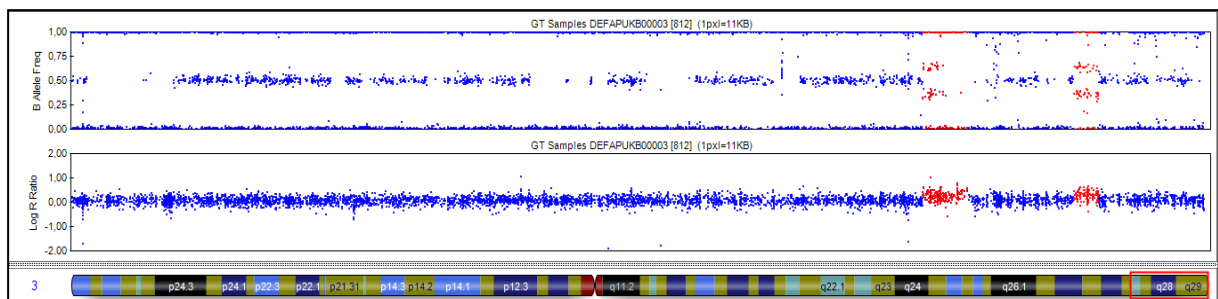
#### 3.11.2. CNV calling

**QuantiSNP calculation:** QuantiSNP (version 1.1 and 2.2, <https://sites.google.com/site/quantisnp/>) was used for CNV detection. Introduced by Colella et al. (2007), QuantiSNP is a computational tool for the detection of copy number variation from BeadArray™ SNP genotype data. It applies a Hidden Markov Model (HMM) consisting of hidden and observed states that represent the unknown copy number of probes in a sample and their normalized intensity measurements in the array. A Bayes Factor provides a probability measure of the strength of evidence, based on the data, for the presence of a copy number variant in a region versus the hypothesis that there is no variant (i.e. that the called CNV is an artifact). The higher the log Bayes factor is, the more likely the CNV is a true positive. In contrast, a CNV with a low log Bayes factor tends to be a false positive. The recommended threshold for true versus false positive is  $BF = 30$ . For the QuantiSNP calculation, two specific values are

of importance; allelic intensity ratio (B-allele frequency, BAF) and signal intensities (log R ratio, LRR).



**Figure 3.4.** Illustration of B Allele frequency and Log R ratio of a deletion CNV. The BAF shows loss of heterozygosity and LRR shows the intensity lower than 0.



**Figure 3.5.** Illustration of BAF and LRR of a duplication CNV.

**B-allele frequency (BAF):** The B-allele frequency (BAF) provides information on the genotype of each SNP marker, depending on whether a red (Cy5) or green (Cy3) fluorescence-labeled nucleotide has been incorporated during the single-base extension on each bead oligonucleotide. If the genotype is homozygous B/B or hemizygous B/-, the BAF is  $\sim 1$  whereas if the genotype is A/A or A/-, the BAF is  $\sim 0$ , and if the scanner detects a mixed color, the genotype is a heterozygote A/B and the BAF is  $\sim 0.5$ . For a heterozygous deletion (only 1 allele present), the BAF will be either  $\sim 1$  or  $\sim 0$ , which corresponds to homozygosity (Figure 3.4). For a duplication (3 copies), the possible homozygous genotypes are BBB or AAA, for which the BAF will be 1 or 0, respectively. Other possible genotypes are BBA or AAB, which lead to BAF values of  $\sim 0.67$  and  $\sim 0.33$ , respectively (Figure 3.5). Homozygous deletions result in a failure of the BAF to cluster. Thus, the BAF may be used to accurately assign copy numbers from 0 to 4 in diploid regions of the genome (Cooper et al. 2008). It also allows detection of copy-neutral events such as segmental uniparental disomy (segmental UPD) or whole-chromosome UPD and identity by descent (IBD), which occurs when a segment of one chromosome is replaced by the other allele without a change in copy number (Peiffer et al. 2006).

Some markers are not SNP markers but CNV markers. The latter give no information about the BAF because they are not polymorphisms. The BAF of these markers is always as if a



homozygous allele were present. Since the BAF alone is not able to unequivocally identify copy number changes, a second value is necessary to evaluate the copy number.

**Log R ratio (LRR):** The Log R ratio (LRR) is a measure of total fluorescence signal intensity measured at every probe. The LRR value for each SNP is calculated as  $LRR = \log_2(R_{\text{subject}}/R_{\text{expected}})$ , in which  $R_{\text{expected}}$  is computed through linear interpolation of canonical genotype clusters obtained from a set of reference samples (Wang and Bucan 2008). The log normalized scale ranges from -1 to 1 and is proportional to the copy number at that locus. The  $\log_2$  of normal copy number (diploid) is 0; if the LRR is below 0, the intensity of the subject is lower than expected. In contrast, the LRR increases with a gain of copy number. The standard deviation of the LRR should not be above 0.30 so unacceptable background noise for reliable CNV detection can be avoided.

## 3.12. CNV analysis

### 3.12.1. Known candidate gene survey

Before performing filtering steps, all called CNVs were checked whether any of the patients carry a deletion or a duplication in known polyposis genes or other CRC related genes (Table 3.4) Genomic positions are based on *Ensembl* genome Browser release 54 (NCBI build 36, hg 18).

### 3.12.2. Filtering CNVs

Two different types of filters were applied: 1) those that are technical in nature to minimize the number of false-positive CNVs (section 3.12.2.1 – 3.12.2.5), and 2) filters that are based on the etiological relevance of the CNV (section 3.12.2.6 – 3.12.2.8). The latter refers to the preferred disease model we considered. This thesis focussed on highly penetrant CNVs, with both dominant and recessive models of inheritance.

#### 3.12.2.1. CNV length

The DGV reports that 95% of CNVs are shorter than 100 kb. Since the majority of CNVs are 1-10 kb in length and tend to be benign CNVs, variants in the 10-50 kb range are considered more important for research (Pinto et al. 2011). Therefore, this thesis set the minimum size for inclusion at  $\geq 10$  kb for CNV length (Engels et al. 2009). Nevertheless, the minimum size for duplication inclusion was set at 20 kb to reduce the number of false positive.

**Table 3.4.** Established causative genes or candidate genes for (hereditary) colorectal tumors

Gene	Chro	Start position	Stop position	Disease/pathway/susceptibility
<i>APC</i>	5q22.2	112101483	112209834	Familial adenomatous polyposis (FAP)
<i>APC2</i>	19p13.3	1401148	1424243	Wnt signalling pathway, depletion of intracellular beta-catenin
<i>AXIN1</i>	16p13.3	277441	342465	Regulator of canonical Wnt signaling pathway
<i>AXIN2</i>	17q24	60955145	60988202	Regulator of beta-catenin pathway
<i>BMPR1A</i>	10q22.3	88506376	88674925	Juvenile polyposis syndrome
<i>BRAF</i>	7q34	140080751	140271033	Known oncogene, somatic mutations in CRC
<i>BUB1B</i>	15q15	38240530	38300627	Gastrointestinal adenomas and carcinomas
<i>CDH1</i>	16q22.1	67328696	67426943	Hereditary diffuse gastric cancer
<i>CHEK2</i>	22q12.1	27413731	27467822	Susceptibility gene for HNPCC
<i>CTNNB1</i>	3p21	41211405	41256943	Central protein of canonical Wnt signaling pathway
<i>EGFR</i>	7p12	55054219	55242524	Known oncogene
<i>EPCAM</i>	2p21	47425801	47467661	Lynch syndrome/HNPCC
<i>FAM123B</i>	Xq11.2	63321722	63342349	Regulator of canonical Wnt signaling pathway
<i>FBXW7</i>	4q31.3	153461860	153675622	Frequently mutated in CRC and colorectal adenomas, Cyclin E degradation
<i>GREM1</i>	15q13.3	30797497	30814158	Member of BMP pathway, susceptible gene for CRC
<i>GSK3B</i>	3q13.3	121028238	121295954	Involved in canonical Wnt signaling pathway, interact with <i>APC</i> , beta-catenin
<i>KRAS</i>	12p12.1	25249449	25295121	Oncogene, frequently mutated in CRC
<i>MAP2K4</i>	17p12	11864866	11987865	Frequently mutated in CRC, MAPK pathway
<i>MLH1</i>	3q21.3	37009845	37067341	Lynch syndrome/HNPCC
<i>MSH2</i>	2p21	47483710	47563871	Lynch syndrome/HNPCC
<i>MSH6</i>	2p16	47863725	47887596	Lynch syndrome/HNPCC
<i>MUTYH</i>	1p32-34	45567501	45578729	<i>MUTYH</i> -associated polyposis (MAP)
<i>NRAS</i>	1p13.2	115051108	115061038	Frequently mutated in CRC
<i>PDGFRA</i>	4q12	54790204	54859168	Frequently mutated in CRC
<i>PIK3CA</i>	3q26.3	180349005	180435189	Cowden syndrome, frequently mutated in CRC
<i>PMS2</i>	7p22.2	5979396	6015263	Lynch syndrome/HNPCC
<i>POLD1</i>	19q13.3	55579420	55613082	Candidate CRC gene for hereditary multiple adenomas
<i>POLD2</i>	7p13	44120811	44129655	Candidate for hereditary multiple adenomas/CRC
<i>POLD3</i>	11q14	73981277	74031413	Candidate for hereditary multiple adenomas/CRC
<i>POLD4</i>	11q13	66875597	66877593	Candidate for hereditary multiple adenomas/CRC
<i>POLE</i>	12q24.3	131710421	131774018	Hereditary multiple adenomas/CRC
<i>POLE2</i>	14q21	49180028	49224685	Candidate for hereditary multiple adenomas/CRC
<i>POLE3</i>	9q33	115209342	115212773	Candidate for hereditary multiple adenomas/CRC
<i>POLE4</i>	2p12	75039283	75050366	Candidate for hereditary multiple adenomas/CRC
<i>PPP2R1B</i>	11q23	111102842	111142379	MAPK pathway, candidate TSG, frequently mutated in CRC
<i>PTEN</i>	10q23.31	89613175	89718511	Cowden's syndrome
<i>PTPRJ</i>	11p11.2	47958686	48148969	Candidate CRC gene

Gene	Chro	Start position	Stop position	Disease/pathway/susceptibility
<i>SFRP1</i>	8p11.21	41238640	41286149	Modulator of Wnt signaling pathway, epigenetic loss in early colorectal adenomas and CRC
<i>SMAD2</i>	18q21.1	43613464	43711510	TGF-beta signaling pathway, frequently mutated in CRC
<i>SMAD4</i>	18q21.1	46810581	46865409	Juvenile polyposis syndrome
<i>SOX9</i>	17q24.3	67628756	67634147	Frequently mutated in CRC
<i>STK11</i>	19p13.3	1156798	1179434	Peutz-Jeghers syndrome
<i>TCF7L2</i>	10q25.3	114700201	114916073	Frequently mutated in CRC, target transcription of Wnt signaling pathway
<i>TP53</i>	17p13.1	7512445	7531642	Li-Fraumeni syndrome
<i>UBC</i>	12q24.3	123962147	123965530	Regulation of various cell signaling pathways by ubiquitination
<i>WIF1</i>	12q14.3	63730674	63801383	Inhibitor of Wnt signaling pathway

### 3.12.2.2. Number of markers

To reduce the signal-to-noise ratio, as had been done in previous studies, the number of consecutive probes was set at  $\geq 5$  for a deletion (Engels et al. 2009; Pinto et al. 2011) and at  $\geq 7$  for a duplication (Venkatachalam et al. 2011).

### 3.12.2.3. Log Bayes Factor

The Log Bayes Factor represents the confidence for a CNV. It is calculated by the QuantiSNP software based on LRR and BAF, where higher values indicate higher statistical reliability. The percentage of false positive CNVs with Log BF lower than 20 is higher than 60% (Priebe et al. 2012), thus minimizes the number of false positive CNV calls; the critical Log Bayes Factor for a deletion and duplication was set at  $\geq 20$  and  $\geq 30$ , respectively.

### 3.12.2.4. Segmental duplications

Segmental duplications are genomic regions with high sequence identity (greater than or equal to 90%) to more than one genomic locus and have been mapped in the human genome. Segmental duplications make up approximately 5% of the human genome. They can appear on the same or different chromosomes (Bailey et al. 2002). Copy number changes mediated by segmental duplications may be benign in the human population (Itsara et al. 2009). A segmental duplication represents an obstacle for the design of specific primers for validating a CNV by qPCR methods. In analogy to Itsara et al. (2010) who excluded segmental duplications and hotspot enrichment from the analysis to avoid ascertainment bias in the results, we also removed putative CNVs in segmental duplication regions.

### 3.12.2.5. Visual inspection on Genome Viewer (Illumina)

The remaining CNVs were inspected for their LRR, BAF, types of probes, and pattern of probes using Genome Viewer software (Illumina). This inspection allowed removing false positive CNVs from the study. A CNV was excluded if all markers are CNV probes, or the majority of probes is located in intronic regions while the last or first probe is located in isolation. An example of a false positive CNV visualized by the Genome Viewer is shown in figure 3.6.

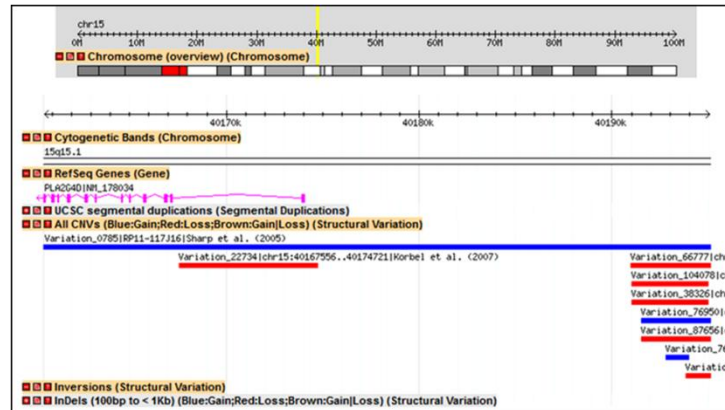


**Figure 3.6.** GenomeViewer illustrating the pattern of a false positive heterozygous deletion called by QuantiSNP. The CNV is located on chr15: 41924943-41935434; length 10492 bp, Max Log BF 26, and involves 17 probes (red dots). All except the last single, right-sided probe are CNV markers, clustering together in a candle-shaped way. The clustered group of CNV probes is located in the intronic region of *WDR76* whereas the exon of the gene (blue circle) is located between the last 2 probes.

### 3.12.2.6. Comparison to the Database of Genomic Variants (DGV)

For overlaps with known CNVs, the Cartagenia Bench™ software set the criteria to filter out common CNVs if there was 100% overlap with variants reported and the number of observed CNVs was  $\geq 3\%$  in the control population. After this in-silico filtering step, the remaining CNVs were manually inspected on the DGV ([http://projects.tcag.ca/ variation/?source= hg18](http://projects.tcag.ca/variation/?source= hg18)). In this thesis, the definition of a common CNV is a CNV which has been reported as a deletion/duplication in at least 1% of study samples in at least two large-scale studies such as those of Redon et al. (2006), Pinto et al. (2007), Jakobsson et al. (2008), Shaikh et al. (2009), and Conrad et al. (2010). The sample sizes of these studies are hundreds to thousands. On the other hand, reports with only a few samples were not used in consideration of classifying a CNV as common or rare. For example, the DGV reports a deletion on chromosome 15:40160584-40195058, shown in figure 3.7 (whole region). Common deletions in the region have been reported by many groups but in all except one report, the deletion is located outside a gene (red bars, right side). A deletion (chr15:40167556-40174721) involving *PLA2G4D*, has been reported in one study only

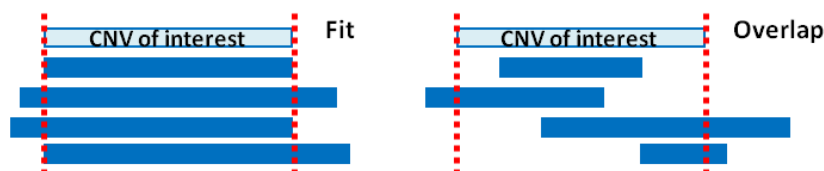
(Korbel et al. 2007) (red bar, left side). The deletion was found in one of two samples. The *PLA2G4D* gene is a reverse strand; the 5'UTR and the first exon were deleted. The duplication (blue bar, whole region), found in one of 47 control samples, has been reported by Sharp et al. (2006).



**Figure 3.7.** Database of Genomic Variants (DGV) reports deletions and duplications on chromosome 15:40160584-40195058. Red bars represent deletions, blue bars represent duplications.

### 3.12.2.7. Comparison to HNR controls

After filtering CNVs with the above criteria, 531 Heinz Nixdorf RECALL (HNR) study controls that were genotyped on the same array and showed log R ratio deviation  $\leq 0.30$  were used for exclusion of common CNVs. CNVs of HNR controls were called with QuantiSNP as had been done in the patient cohort. Called CNVs smaller than 1 kb were excluded from the study because they are defined as ‘indels’, not ‘CNVs’. The patients’ CNVs were compared against the controls’ CNVs and were removed from the study if they were present in more than 1 control individual (frequency  $> 0.2\%$ ) and either partly or completely overlap with those in controls (Figure 3.8). CNVs found more than once in the control cohort were described to be common CNVs.



**Figure 3.8.** Definitions of the position of CNVs used for comparing between patient CNVs (CNVs of interest) and control CNVs (blue bars). Patient CNVs completely covered by control CNVs (left) or partly overlapping the control CNVs (right) were considered to be found in the control cohort.

### 3.12.2.8. Gene content

CNVs were checked against *Ensembl* genome browser 54 to exclude non-genic CNVs and genic CNVs disrupting only an intronic part (containing no coding sequence). Only the CNVs containing protein coding genes were included.

### 3.13. CNV validation

#### 3.13.1. Quantitative PCR (qPCR)

Real-Time quantitative polymerase chain reaction (qPCR) is similar to a simple PCR except for its ability to monitor the progress of the PCR as it occurs in real time. The PCR product is measured after each round of amplification while with traditional PCR, the amount of PCR product is measured only at the end of amplification. The amplified product is measured based on fluorescent label. Two common methods for the detection of products in quantitative PCR are 1) non-specific fluorescent dyes that intercalate with any double-stranded DNA and 2) sequence-specific DNA probes consisting of oligonucleotides that are labeled with a fluorescent reporter which permits detection only after hybridization of the probe with its complementary sequence. They all link the amplification of DNA to the generation of fluorescence which can simply be detected with a camera during each PCR cycle. Hence, as the number of gene copies increases during the reaction, so increases the fluorescence.

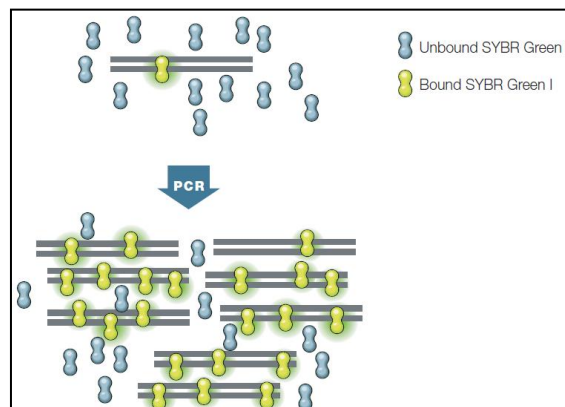
Quantitative PCR can be used to quantify nucleic acids by two common methods: relative quantification and absolute quantification. Relative quantification is based on internal reference genes to determine fold-differences in expression of the target gene. The quantification is expressed as the change in expression levels of mRNA interpreted as complementary DNA. Absolute quantification gives the exact number of target DNA molecules by comparison with DNA standards using a calibration curve. It is therefore essential that the PCR of the sample and the standard have the same amplification efficiency. Relative quantification is easier to carry out than absolute quantification as it does not require a calibration curve: the amount of the studied gene is compared to the amount of a control housekeeping gene.

#### 3.13.2. CNV validation by qPCR using SYBR Green I

SYBR Green I is a DNA-intercalating (DNA minor-groove binding) dye, which is able to nonspecifically bind to double stranded DNA to detect a PCR product during PCR cycles. SYBR Green I exhibits little fluorescence when it is free in solution, but its fluorescence intensity increases up to 1,000-fold when it binds double-stranded DNA (Figure 3.9).

To validate the copy number of putative CNVs, qPCR using SYBR Green was performed on an ABI Prism 7900HT Fast Real-Time PCR System (Life Technologies) with exact amounts of DNA samples. For each CNV, three primer pairs were designed using the online program Primer3. The primers were put in coding regions (if possible) at the beginning, in the middle, and at the end of each CNV. All primer sequences are presented in Tables A4 and A5. Three housekeeping genes (*BNC1*, *CFTR*, *RPP38*) were used as internal controls for normalization (Table A6). The reason for using housekeeping genes is to correct for non-specific variation.

Four anonymous healthy controls were used for assessment of the variability in copy number. Each assay was run in triplicate and contained a no template control (NTC).



**Figure 3.9.** Illustration of the SYBR Green I binding to double-stranded DNA in real-time PCR. The fluorescence signal is increased in proportion to the amount of the target DNA amplified. (Figure adapted from [www.gene-quantification.de/real-time-pcr-guide-bio-rad.pdf](http://www.gene-quantification.de/real-time-pcr-guide-bio-rad.pdf))

### Reaction components

The Power SYBR® Green I PCR Master Mix (Life Technologies) is supplied in 2X concentration. The mix contains SYBR Green I Dye, AmpliTaq Gold® DNA Polymerase, dNTP, passive reference (ROX™ dye), and optimized buffer components. The passive reference provides an internal reference for data normalization, which is necessary to correct well-to-well fluorescence fluctuations. In total 10 µl of each reaction contains 20 ng of template DNA. Reagents for each reaction were provided as shown in table 3.5.

**Table 3.5.** Amount and final concentration of qPCR reagents per reaction

Reagent	Volume (µl)	Final concentration
2X Power SYBR® Green I PCR Master Mix	5	1X
For primer [10 pmol/µl]	0.2	0.2 µM/reaction
Rev primer [10pmol/µl]	0.2	0.2 µM/reaction
Genomic DNA [10ng/µl]	2	20 ng/reaction
H <sub>2</sub> O	2.6	
Total	10	

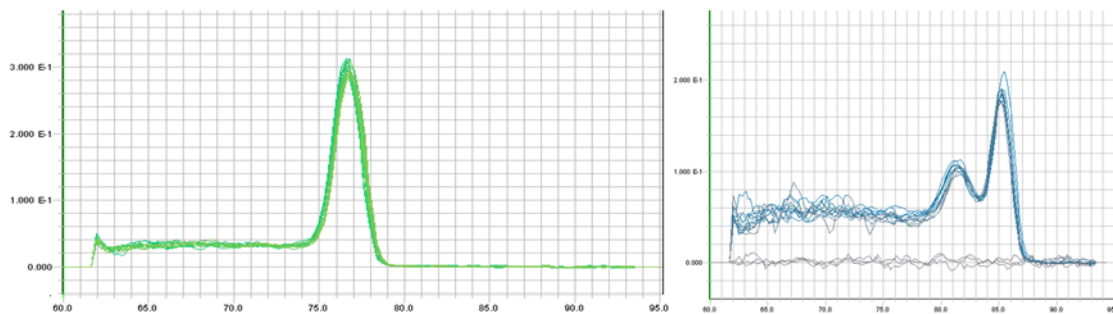
### Cycling step

The conditions for amplifications were started with a temperature of 50°C for 2 minutes, then denaturation at 95°C for 10 minutes, followed by 40 cycles of 95°C for 15 seconds and a combined annealing and extension step at 60°C for 60 seconds, followed by melting temperature curve analysis.

### 3.13.3. Data analysis

Absolute quantification was used to determine the absolute copy number. The dissociation stage was performed to recognize the presence of unspecific product because SYBR GreenI binds to any double-stranded DNA including nonspecific double-stranded DNA sequences, which leads to false positive signals. A multi-peak might be caused by unspecific products or an unspecific primer elongation or primer dimer or SNPs (Figure 3.10). In this case, primers were redesigned.

The signals from the PCR amplification were detected by Sequence Detection Software (SDS) version 2.2.2 (Life Technologies). Results were expressed in terms of the threshold cycle (Ct) value, which was calculated with SDS 2.2.2 using standard settings for Ct and autobaseline (Figure 3.11).



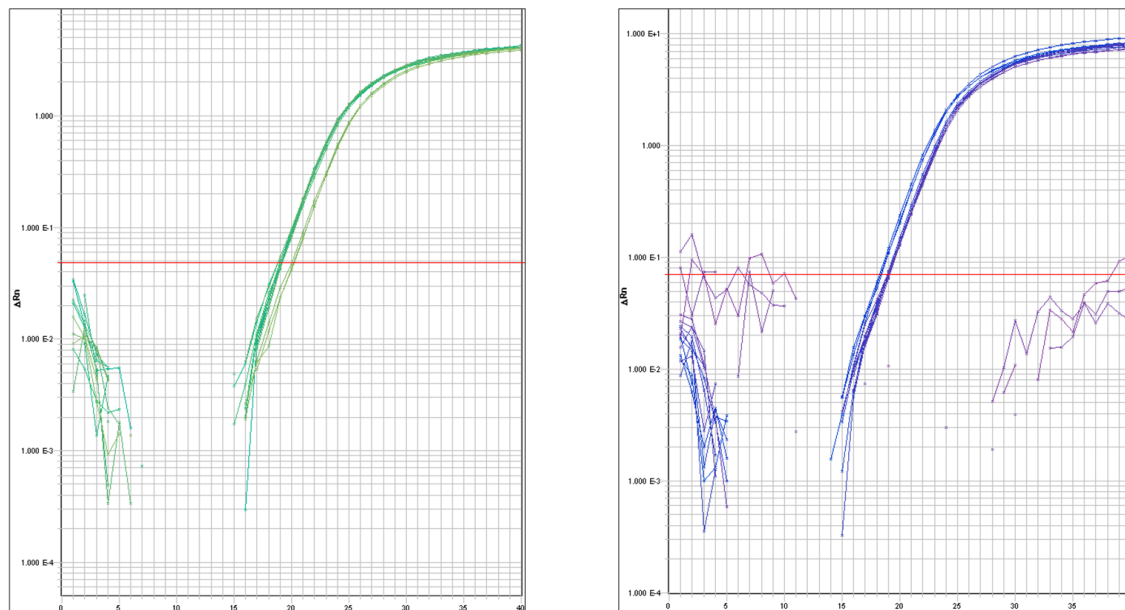
**Figure 3.10.** The dissociation curves represent the specific product (left) and the unspecific product (right) which leads to a false Ct result. The X axis represents temperature (°C) and the Y axis represents derivative of log.

### 3.13.4. Copy number calculation ( $2^{-\Delta\Delta C_t}$ method)

Since the assays were done in triplicate, an average of the Ct was calculated in all assays. The standard deviation (SD) must not be higher than 0.20 to ensure the reliability of the Ct value. Assays with SD higher than 0.20 were repeated.

The delta delta Ct ( $2^{-\Delta\Delta C_t}$ ) method was used to identify the copy number. This method directly compares the Ct values between patients and controls. The first step is normalization and  $\Delta C_t$  was generated by subtracting the Ct of the target from the Ct of the reference ( $\Delta C_t = C_t \text{ CNV} - C_t \text{ Ref Gene}$ ). Then the  $\Delta C_t$  of the case was compared to the  $\Delta C_t$  of the control ( $\Delta\Delta C_t = \Delta C_t \text{ patient} - \Delta C_t \text{ control}$ ). The copy number was then calculated with the formula;  $CN = 2 \times 2^{-\Delta\Delta C_t}$  (Livak and Schmittgen 2001).





**Figure 3.11.** The amplification plots show (left) the difference between the Ct value of a PCR product with a normal copy number (dark green) and the Ct value of a PCR product with loss of a copy number (light green), and (right) the difference of the Ct value between the normal copy number (dark blue) and the Ct value of a gain of copy number (blue). The x-axis represents the cycle of the PCR. The y-axis represents the normalization of the fluorescence signal between the reporter signal and the baseline signal ( $\Delta Rn$ ). The red line is the threshold at which the reaction reaches fluorescence intensity above background.

### 3.14. Co-segregation analysis

To see whether or not the CNV segregates with the phenotype of family members, and based on whether the DNA of patients' family members is available, the copy number was evaluated by qPCR with the same primer set as was used for validation of the CNV in the probands. qPCR (see section 3.13) was performed to check whether the CNV of the index patient was also present in affected relatives or absent in non-affected relatives.

### 3.15. Gene expression analysis

#### 3.15.1. Gene expression in human colon cDNA

To check whether the identified candidate genes are known to be expressed in colon tissue, two publicly available databases were used: the EST profiles reported in the *UniGene* database ([www.ncbi.nlm.nih.gov/unigene](http://www.ncbi.nlm.nih.gov/unigene)), and RNA expression data from normal human tissues reported in *GeneCards* ([www.genecards.org](http://www.genecards.org)). Genes were considered to be expressed if the value of transcripts per million (TPM) was  $> 0$ . To exclude false negative results in the accessed databases, the expression of candidate genes with reported non-

expression in human colon tissue was examined using commercial first strand cDNA of human adult colon mucosa (Amsbio). Genes which are unexpressed in human colon cDNA were removed from the list of candidates.

### 3.15.2. PCR and agarose gel electrophoresis

PCR Ready First Strand cDNA Human adult normal colon tissue (Amsbio) [2.5 ng/ $\mu$ l] was used for this study. REDTaq ReadyMix PCR Reaction Mix with MgCl<sub>2</sub> (Sigma Alrich) was used for semi-quantitative PCR (Pasternack et al. 2008). A primer pair for each gene was designed and synthesized as described in section 3.7. Amplicon length was set at 400-600 bp and, if possible, primers were put in exon-exon boundaries and the exon must be present in every isoform when there is more than one isoform. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as an endogenous control, and positive control cDNA for the analysis was taken from Human MTC™ Panels I & II (Clontech) (placenta, liver, stomach, esophagus, and testis), hair follicle, and peripheral blood leukocytes. A new primer pair was designed to repeat the test in case no band showed up on the gel and, in particular, no positive tissue was available.

#### PCR reaction components

Before preparing the reaction, primers were diluted 1:10 in H<sub>2</sub>O to a working concentration of 10  $\mu$ M. Reagents used for PCR reaction preparations are described in table 3.6. The PCR cycle is described in section 3.7.

The PCR product was checked with a 2% agarose gel electrophoresis. If PCR products obtained on cDNA showed a visible band on the control tissue (e.g. placenta or testis) but not on the colon tissue, the primer was assumed to work well but the candidate gene was not expressed in colon. If the product is not present on any tissue, the gene is either not expressed in both, control tissues and colon tissue, or the primer is not working. In the latter scenario, a new primer pair was designed and the test was repeated to confirm the result.

**Table 3.6.** Reagents used for the PCR reactions

Reagents	Volume ( $\mu$ l)	Final concentration
2X REDTaq	12.5	1X
10 $\mu$ M Forward primer	1	0.4 $\mu$ M
10 $\mu$ M Reverse primer	1	0.4 $\mu$ M
[2.5 ng/ $\mu$ l]Template cDNA	1	2.5 ng/reaction
H <sub>2</sub> O	9.5	

### 3.16. Network analysis

To further prioritize our candidate genes, a network and pathway analysis was performed in collaboration with Prof. Dr. Holger Fröhlich at the Bonn-Aachen International Center for Information Technology in Bonn. The network analysis was performed twice. For the first analysis, protein-protein interaction was performed to compare our candidate genes and established polyposis genes, published candidate genes for colorectal cancer, and putative disease genes from a polyposis GWAS (unpublished data). The second analysis was performed via the *Steiner Tree* algorithm (Sadeghi and Frohlich 2013). Network analysis, pathway analysis, and enrichment analysis were carried out according to KEGG pathway and GO terms (biological process) with the conditional hyper-geometric test approach (Falcon and Gentleman 2007) in both CNV candidate genes and known candidate genes. All human genes were employed as statistical background.

The pathway analysis of 180 candidate genes was also performed by an online Ingenuity pathway analysis tool ([www.ingenuity.com/products/ipa](http://www.ingenuity.com/products/ipa)). This online tool evaluates the set of input genes as to whether there are known relationships between the genes and biofunctions with subcategory of diseases, and cellular functions in the Ingenuity Knowledge Base. The probability that each biological function and/or disease assigned to that data set is due to statistical chance was calculated by a right-tailed Fischer's Exact Test. Overrepresentation of the molecules in a given process was considered to be statistically significant when  $P < 0.05$ . The over-represented functional or pathway processes are those with more focus molecules than expected by chance (Mitra et al. 2012).

### 3.17. Candidate gene prioritization

To select the most promising candidate genes for further work-up, many conditions and criteria were applied to prioritize candidate genes as described below.

#### 3.17.1. Frequency of finding

Candidate genes present in more than one patient were included for further study and prioritized as being most interesting genes.

#### 3.17.2. Segregation analysis

When DNA from any family member of the study patients was available, the copy number of the candidate CNV region identified in the study patient was examined in the relatives to prove whether the CNV co-segregates with the phenotype or not. If the CNV co-segregated with the phenotype, the genes in the CNV would be further studied and ranked as top

prioritization. If the CNV did not segregate with the phenotype, the CNV might not be causative or is assumed to have a low to moderate penetrance.

### 3.17.3. Data mining

Another approach we used to rank our candidates was text mining by access through relevant databases and literature of individual genes to maximize the chance of identifying relevant genes (Moreau and Tranchevent 2012). Many online databases such as COSMIC (Catalogue of somatic mutations in cancer), DAVID (Database for Annotation, Visualization and Integrated Discovery), GENATLAS (<http://genatlas.medecine.univ-paris5.fr/>), GeneCards ([www.genecards.org/](http://www.genecards.org/)), Gene Codis (Gene annotations co-occurrence discovery), the GWAS catalog (A Catalog of Published Genome-Wide Associations Studies), Human Gene Mutation Database (HGMD), the KEGG database (Kyoto Encyclopedia of Genes and Genomes), OMIM (Online Mendelian Inheritance in Man), PUBMED (<http://pubmed.com/>), and STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) were used for compiling all relevant information of the candidate genes including functions, pathways, related phenotypes, gene interactions, tissue specification, GWA studies, and publications.

## 3.18. TaqMan® gene expression analysis

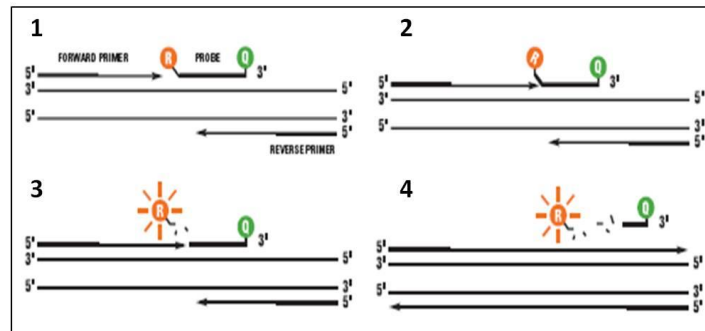
### 3.18.1. Basic principle

The TaqMan®-based detection probe uses a fluorogenic probe and represents a more specific quantitative detection method than SYBR Green I. The latter is an intercalator dye, which is able to detect both specific and non-specific products, whereas fluorogenic probes amplify only a specific product and enable the development of a real-time detection method.

The probes are dual labeled; a reporter fluorescence dye (VIC or FAM) labels on the 5' end and a non-fluorescent quencher (NFQ) and minor groove binder (MGB) are at the 3' end of the probe. The fluorophore is a molecule that emits light of a certain wavelength after having first absorbed light of a specific but shorter wavelength. The quencher is a molecule that accepts energy from a fluorophore in the form of light and dissipates this energy in the form of light. The MGBs increase the melting temperature ( $T_m$ ) without increasing probe length.

The fluorophore is excited by the machine and passes its energy via FRET (Fluorescence Resonance Energy Transfer) to the quencher (Cardullo et al. 1988). If the target sequence is present, the probe anneals downstream from one of the primer sites and is cleaved by the 5' nuclease activity of Taq DNA polymerase as this primer is extended. Additional reporter dye molecules are cleaved from their respective probes with each cycle, resulting in an increase in fluorescence intensity proportional to the amount of amplicon produced (Figure 3.12). The

increase in fluorescence occurs only if the target sequence is complementary to the probe and is amplified during PCR. Therefore, without specific amplification a sequence is not detected.



**Figure 3.12.** Schematic of the FRET method. **1)** A fluorescent reporter (R) dye and a quencher (Q) are attached to the 5' and 3' ends of a TaqMan probe respectively. **2)** The reporter dye emission is quenched. **3)** The DNA polymerase cleaves the reporter dye during each PCR cycle. And **4)** the reporter dye separates from the quencher and emits its fluorescence. (Figure adapted from [www.appliedbiosystems.com/absite/us/en/home/applications-technologies/real-time-pcr/taqman-and-sybr-green-chemistries.html](http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/real-time-pcr/taqman-and-sybr-green-chemistries.html))

### 3.18.2. Relative quantitative PCR (RT-PCR)

Duplex real-time PCR allows the amplification of two target sequences in a single reaction. Two assays must have a different specific probe labeled with a unique fluorescent dye, resulting in different observed colors for each assay. Typically one probe is used to detect the target gene, another probe is used to detect an endogenous control. The endogenous control is necessary for normalization of each sample to adjust for differences in total DNA content.

This study was performed to quantify the expression of *CTNNB1* or *MUTYH*. TaqMan® gene expression assays for this study were designed and generated by Life Technologies. Each assay was composed of specific primers and probes for *CTNNB1* or *MUTYH*. Human Cyclophilin and Human Beta-2-microglobulin (huβ2M) (Life Technologies) were used as endogenous controls. The assays for reference genes were labeled with VIC while the assays for the target genes were labeled with FAM. Ten anonymous healthy controls (5 males, 5 females) were used for the statistical comparison.

Each assay was run in triplicate. Each run contained a no-template control (NTC). For valid Ct values, the standard deviation (SD) of each assay must not be higher than 0.20.

### 3.18.3. Reaction components

Reagents, volume and final concentrations for a duplex qPCR reaction are presented in table 3.7. *CTNNB1* assay was combined with huβ2M while *MUTYH* was combined with the

Cyclophilin A assay. Exact pipetting is a crucial requirement for this experiment as even the smallest change in the amount of template could lead to wrong results.

**Table 3.7.** Reagents used for TaqMan gene expression study

Reagent	Volume (μl)	Final concentration
TaqMan® Gene Expression Master Mix (2X)	5	1X
TaqMan® Gene Expression Assay (20X FAM)	0.5	1X
TaqMan® Endogenous Control (20X VIC)	0.5	1X
cDNA (50 ng RNA equivalent/μl)	2	100 ng RNA equivalent/reaction
RNase-free water	2	
Total	10	

#### 3.18.4. Cycling step

The qPCR was performed on an ABI Prism 7900HT Fast Real-Time PCR System (Life Technologies) with standard amplification programs as for qPCR using SYBR Green I. The amplifications were started at 50°C for 2 minutes, then denaturation at 95°C for 10 minutes, followed by 40 cycles at 95°C for 15 seconds, and a combined annealing and extension step at 60°C for 60 seconds. Melting temperature curve analysis is not needed for TaqMan assays.

#### 3.18.5. Data analysis

The Comparative  $C_T$  ( $\Delta\Delta C_T$ ) method (relative quantification) was used to analyze changes in gene expression in a target sample relative to a reference gene. The original amount of transcript was then obtained from that normalized  $C_t$  value by calculating  $1/(2^{x C_t})$ . A reasonable value of the result was obtained by multiplying it with an arbitrary factor of 100,000,000.

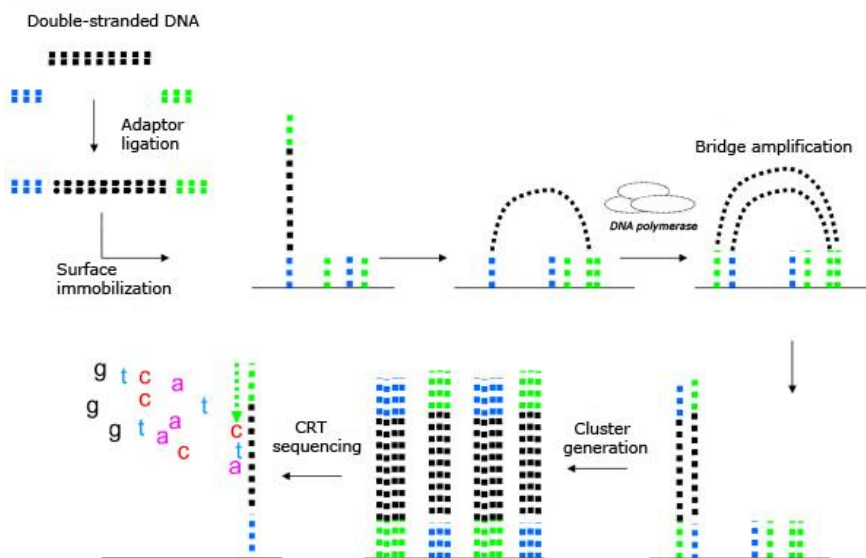
To determine differences in gene expression level between a patient and controls, statistical significance was analyzed by R software version 3.0.2 using a two-sided Wilcoxon test. The Wilcoxon test is a non-parametric statistical test to compare two independent samples. The statistical difference was considered significant when the *p-value* was less than 0.05.

### 3.19. Targeted next generation sequencing

#### 3.19.1. Basic principle

Next generation sequencing (NGS) refers to high throughput sequencing technologies that have emerged during the last decade. Each platform is different in its methods of clonal amplification of short DNA fragments as a genomic library template and how these fragment libraries are subsequently sequenced through repetitive cycles to provide a nucleotide readout. The sequencing is conducted in either a stepwise iterative process or in a continuous real-time manner. By virtue of the highly parallel process, each clonal template is sequenced individually and can be counted among the total sequences generated (Myllykangas et al. 2011). In principle, there are three major steps: sample preparation, sequencing by collection of images, and data analysis.

The sample preparation steps generally involve random breakage of genomic DNA and product filtering into a sub-pool of suitable fragment sizes (50–400 bases). Filtered fragments are further processed into a library by applying adapters, through ligating containing a universal primer motif, to the ends of each fragment. A successful amplification of the DNA fragments results in one clonal product per bead, or a cluster in case using a glass slide. Beads and clusters produce an amplified signal, representing the average of the original molecule during the sequencing step (Figure 3.13).



**Figure 3.13.** Illumina sample procedure. Double strand DNA is fragmented and ligated with an adaptor, then clusters are generated by bridge amplification, and sequencing is performed during synthesis (Figure adapted from Rizzi et al. (2012)).

The development of algorithms and software that are able to determine the success of the experiment and turn data into manageable results is crucial for interpreting the increasing piles of generated sequences. In principle, the NGS analysis pipeline consists of 4 major

steps; 1) read cleaning or raw data analysis: this first step is to discard failed reads and collect only cleaned reads; 2) read mapping: this step is to align the reads against a reference genome; 3) variant calling: to detect genetic variations such as SNPs, deletions/insertions, and CNVs; 4) variant annotation: this step is linking the variants to biological information, i.e. to specific genes and transcripts to determine the functional consequences such as the mutation type, and to appropriate databases.

### **3.19.2. Library preparation, target enrichment, and sequencing**

In this study, the targeted NGS was performed in collaboration with the *Cologne Center for Genomics (CCG)*. 192 DNA samples (validation cohort) were enriched with the TruSeq® Custom Enrichment kit (Illumina). The oligonucleotide probes were designed by means of the Illumina DesignStudio (<http://designstudio.illumina.com/>). One µg genomic DNA extracted from leukocytes with standard protocols was fragmented using sonication technology (Bioruptor) and fragments were end repaired and adapter ligated using the TruSeq® DNA HT Sample Preparation Kit (Illumina). Custom capture of targeted regions was performed on pools of 12 indexed libraries with the TruSeq enrichment protocol (Illumina). The captured DNA was sequenced by an Illumina HiSeq2000 sequencer with 2x100bp paired-end reads achieving 30x coverage for at least 96% of targeted bases. Data were filtered using Illumina Realtime Analysis® (RTA) software.

### **3.19.3. Alignment, genotype calling, and variant annotation**

Primary data were filtered according to signal purity using the Illumina Realtime Analysis (RTA) software version 1.8. Subsequently, the reads were mapped to the human genome reference build GRCh37 with the *BWA* version alignment algorithm (Li and Durbin 2009). GATK version 1.6 (McKenna et al. 2010) was used to remove duplicated reads, perform local realignment around known indels from the 1000 Genomes Pilot, and recalibrate base quality scores. Variant calling was performed using SAMtools version 0.1.7 (Li et al. 2009) for InDel detection. Scripts developed in-house at the CCG were applied to detect point mutations and overlaps with known variants. The sequences were stored and further analyzed in a user-friendly database (VARBANK, version 2.6) developed by CCG, Germany (<http://varbank.ccg.uni-koeln.de/>).

### **3.19.4. Data analysis and filter**

The annotated sequences were processed through the VARBANK pipeline to detect protein changes, affected donor and acceptor splice sites, and overlaps with known variants. Acceptor and donor splice site mutations were analyzed with a Maximum Entropy mode (Yeo and Burge 2004), and filtered for effect changes. In particular, filtering was performed for high-quality, rare (MAF < 0.01) autosomal variants using allele frequencies from the 1000



*Genomes* database and the *Exome Variant Server*. Filtering was also performed against an inhouse-database containing variants from 511 exomes from epilepsy patients in order to exclude pipeline-related artifacts ( $MAF < 0.02$ ). The cDNA bases were numbered according to the gene reference sequence in *GenBank*, where 1 corresponds to the A of the ATG translation initiation codon. All relevant information of the pre-filtered variants was downloaded as an excel sheet and afterwards further filtered and analysed.

### 3.19.5. Validation of results

All variants were visually inspected with the VARBANK Read Browser to exclude obvious false positive variants and artifacts. Afterwards, only rare truncating mutations (nonsense mutations, insertions, deletions, and splice site mutations) were confirmed by Sanger sequencing. To evaluate the functional impact of missense and splice site mutations on the structure and function of a protein, four in-silico tools (Polyphen-2, Mutation Taster, SIFT, and BDGP) were applied.

## 3.20. Genotyping based on MassExtend Reaction (Sequenom®)

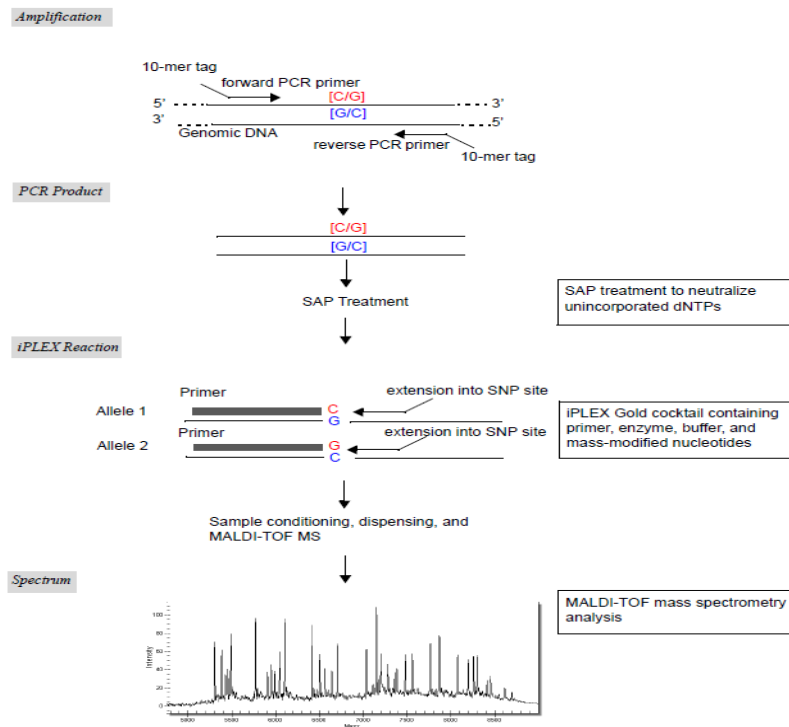
### 3.20.1. Basic principle

The Sequenom MassArray MALDI-TOF (matrix-assisted laser desorption/ionization time-of-flight) system is a compact mass spectrometer. The MALDI-TOF MS has been used for genotyping a limited set of SNPs in a large number of individuals by analysis of the mass of single base extension oligonucleotides specific to the SNP of interest. The Sequenom MassARRAY iPLEX SNP typing platform uses MALDI-TOF MS coupled with single-base extension PCR for high throughput multiplex SNP detection, which is capable of multiplexing up to 40 SNPs per single reaction. The homogeneous assay consists of PCR amplification of the target, followed by incubation with shrimp alkaline phosphatase (SAP) to inactivate unincorporated nucleotides (Figure 3.14). Then a primer is hybridized adjacent to the SNP of interest, extended through the SNP by a single ddNTP under standard thermocycling conditions, followed by reaction termination. The mass of the primer extension product is analyzed and used to determine the sequence of the nucleotide at the SNP site.

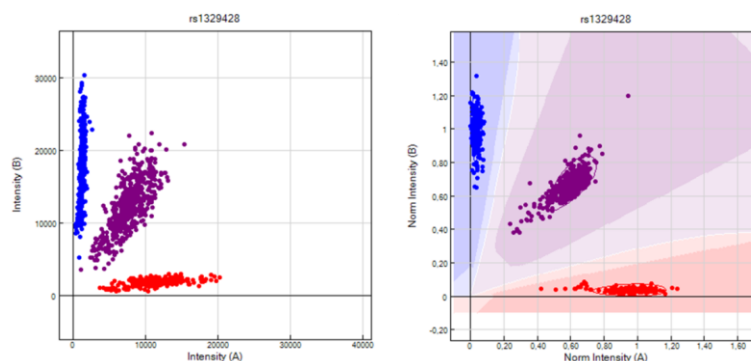
### 3.20.2. Selection of the genotyped SNPs

From a recently performed GWAS in 178 unexplained adenomatous polyposis patients and 536 HNR control individuals based on genotyping data from the above mentioned SNP array experiments, the most promising SNPs were selected for a replication study using the Sequenom platform for genotyping. The most promising SNPs were mainly selected by a top-down approach, i.e. including the SNPs with the highest association according to p-

values. 140 SNPs showed  $p$ -values  $< 10^{-4}$ , and odds ratios of 0.284-4.326. All 140 SNPs showed appropriate clustering on GenomeStudio (Figure 3.15). The number of SNPs for genotyping was reduced from 140 to 119 by exclusion of SNPs which are in strong LD ( $r \geq 0.8$ ) with each other, i.e. which are located on the same haploblock. From all SNPs in LD, only two were included to reduce their number.



**Figure 3.14.** Sequenom® MassARRAY® MALDI-TOF MS SNP genotyping schematic. The target is PCR amplified followed by SAP incubation to inactivate unincorporated nucleotides. During the iPLEX Gold reaction, an internal primer is extended through the SNP by a single ddNTP. The mass of the primer extension product is determined by MALDI-TOF MS to determine the sequence of the nucleotide at the SNP site. (Figure adapted from iPLEX™ Gold Application Guide, Sequenom)



**Figure 3.15.** Cluster plot of SNP rs1329428 showed three genotypes; AA (horizontal, red dots), BB (vertical, blue dots), and AB (diagonal, purple dots). A plot of normalized values (right figure) shows clear genotypes calling.

### 3.20.3. DNA preparation

Ten  $\mu\text{l}$  of DNA (20 ng/ $\mu\text{l}$ ) was prepared and delivered in 96-well plates by our collaborators. Before dilution, the DNAs were randomly measured regarding concentration using a Thermo Scientific NanoDrop™ ND-1000 Spectrophotometer. The DNAs were diluted to 5 ng/ $\mu\text{l}$  as working concentration, then 2  $\mu\text{l}$  of the working DNAs were transferred to a 384-well PCR plate and air-dried overnight before the PCR step. Each PCR plex included one no-template control (NTC) and duplicate DNA samples.

### 3.20.4. Assay and primer design

The Human Genotyping Tools Online ([www.mysequenom.com](http://www.mysequenom.com)) was used to retrieve flanking sequences of all 119 SNPs and to check positions of proxy SNPs, which could lead to a problem when designing primers. Forward and reverse PCR primers as well as extension primers for single base extension were automatically designed. The optimal amplicon size was set at 80 to 120 bp. A 10-mer tag (5'-ACGTTGGATG-3') was added to the 5' end of each PCR primer to avoid confusion in the mass spectrum, and SBE primers were 5' tailed with non-homologous sequences varying in length to create large enough mass differences between the different SBE products to be detected by MALDI-TOF MS.

In this study, the software was not able to design primers for four of the SNPs (rs10148733, rs4236975, rs4967946, and rs11709614); therefore these SNPs were replaced by proxy SNPs. To search for suitable proxy SNPs, the SNP Annotation and Proxy Search (SNAP) online tool (Broad Institute) ([www.broadinstitute.org/mpg/snap/](http://www.broadinstitute.org/mpg/snap/)) was used with the HapMap release 21 SNP dataset. The  $r^2$  threshold was set at 1. Three SNPs, rs10148733, rs4236975, and rs4967946, were replaced by rs60455014, rs4236978, and rs12935619, respectively. Nevertheless, there was no proxy SNP available for rs11709614: thus, this SNP was genotyped with the TaqMan Genotype assay instead (see section 3.21).

To design the multiplex genotype assay, Sequenom's MassARRAY Designer software was used. Each plex is able to contain up to 40 SNPs. The range of weight was set at 4300-9000 Da and the minimum peak separation was set at 30 to reduce ambiguity of peaks. The 118 SNPs were divided into four plex's; information on all SNPs, plexes and primers used in the experiments are given in Table A7.

### 3.20.5. PCR step

The PCR assay pool plexes consisted of the multiplexed forward and reverse PCR oligonucleotide primers for each reaction present together in one multiplexed assay pool. PCR was performed in 5  $\mu\text{l}$  reaction volumes containing 0.5 U of Taq polymerase, 5-10 ng of genomic DNA, 100 nM of PCR primers, and 500  $\mu\text{M}$  of dNTPs.

**PCR cycle**

Step	Temperature (°C)	Duration	Number of cycle
Initial denaturation	95	15 min	1X
Denaturation	95	20 sec	45X
Annealing	56	30 sec	
Extension	72	1 min	
Final extension	72	3 min	1X

**3.20.6. Digestion step**

Shrimp alkaline phosphatase (SAP) dephosphorylates unincorporated dNTPs by cleaving the phosphate groups from the 5' termini. To remove remaining unincorporated dNTPs, PCR products were treated with 0.5 U SAP by incubation at 37°C for 40 minutes, followed by enzyme inactivation by heating at 85°C for 5 minutes. 480 reactions of reagents were prepared for robot pipetting into 384 well-plate.

Reagent	Volume (μl)	480 reactions (μl)
Shrimp Alkaline Phosphatase (SAP) buffer	0.17	81.6
SAP enzyme	0.3	144
H <sub>2</sub> O	1.53	734.4
Total	2	960

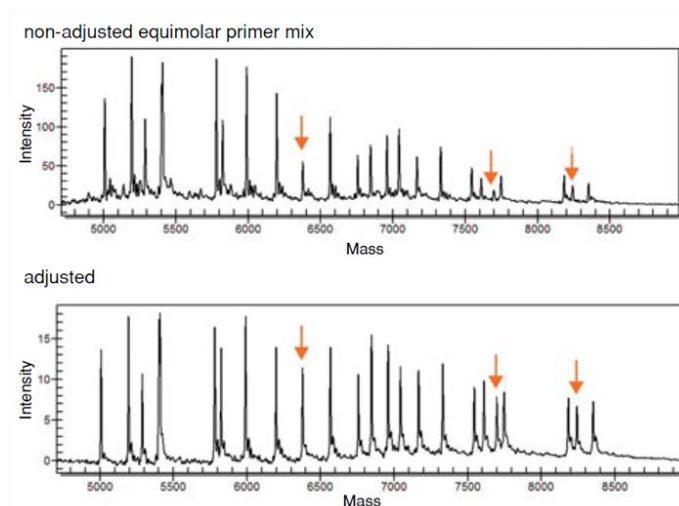
**3.20.7. Extension primer adjustment**

Adjusting the concentrations of oligos to equilibrate signal-to-noise ratios was performed to increase signal-to-noise ratios, which tend to decrease in multiplex experiments. Eventually, the signals become indistinguishable from noise, resulting in calling errors (Figure 3.16). A general method to adjust extension primers is to divide the primers into a low mass group and a high mass group. All primers in the high mass group are doubled in concentration with respect to those in the low mass group.

In this study, mass groups were generated using the primer adjustment tool of the Typer software version 3.4 and 4.0. The primers were sorted by weight into four groups, with the highest mass group diluted to 14 μM, the second and third highest groups to 11.6 μM and 9.3 μM, respectively, and the last group with the lowest mass diluted to 7 μM.

**3.20.8. Extension step**

The primer extension or iPLEX reaction is a method for detecting single-base polymorphisms or small insertion/deletion polymorphisms in amplified DNA.



**Figure 3.16.** Spectra of adjusted and non-adjusted oligos in an assay pool (Figure adapted from Sequenom's iPLEX Gold Application Guide).

Two  $\mu\text{l}$  of an iPLEX Gold extension reaction cocktail, containing extend primer, buffer, enzyme, and mass-modified ddNTPs, was added to the purified PCR products. During the iPLEX reaction, the primer was extended by one mass-modified nucleotide depending on the allele and the design of the assay.

Reagent	Volume ( $\mu\text{l}$ )	480 reactions ( $\mu\text{l}$ )
10X I-PLEX buffer plus	0.2	96
I-PLEX termination mix	0.2	96
UEP primers mix	0.94	451.2
I-PLEX enzyme	0.041	19.68
H <sub>2</sub> O	0.619	297.12
Total	2	960

### Extension step cycle

Step	Temp (°C)	Duration	Number of cycle	
Initial denaturation	94	30 sec	1X	
Denaturation	94	5 sec	1X	40X
Annealing	52	5 sec	5X	
Extension	80	5 sec		
Final extension	72	3 min	1X	

### 3.20.9. Clean up reaction

Each extension product was diluted with 16  $\mu\text{l}$  of water before desalting of the products with 6  $\mu\text{g}$  of CLEAN resin (Sequenom). The plate was turned upside down for 10 minutes to mix-

shake products and resin, then centrifuged at 4000 rpm for 7 minutes before dispensing the cleaned extension product on a chip using a Nanodispenser.

### **3.20.10. Dispensing DNA on a chip**

The MassARRAY Nanodispenser was used to dispense reaction products. Approximately 20 nl of the product was dispensed onto a 384-format SpectroCHIP (Sequenom). Calibrant solution was spotted onto the chip at 10 positions before the chip was transferred to the Mass Spectrometer.

### **3.20.11. Mass spectrometry**

The array was placed into a Mass Spectrometer and each spot was then shot with a laser under vacuum by the MALDI-TOF method. MALDI-TOF MS analysis was performed on a MassARRAY Compact Analyzer (Sequenom). Data acquisition was automatically performed by SpectroAcquire, with ten laser shots per raster position and a threshold of five good spectra per sample pad. Once the sample molecules were vaporized and ionized, they are transferred electrostatically into a time-of-flight mass spectrometer (TOF-MS), where they were separated from the matrix ions, individually detected based on their mass-to-charge ( $m/z$ ) ratios, and analyzed. Detection of an ion at the end of the tube was based on its flight time, which was proportional to the square root of its  $m/z$ .

### **3.20.12. Data analysis**

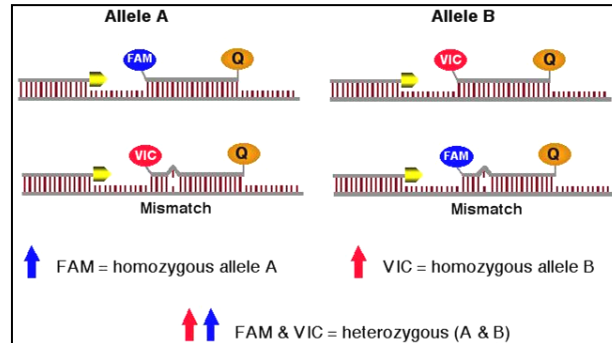
Quality control was monitored by including a duplicate sample and a negative control in each 384-well plate. The data were automatically analyzed with MassARRAY™ Typer Software version 3.4 and then were manually checked. Genotyping calls were viewed in call cluster plots, and peak intensities were reviewed in each respective sample spectrum. Sample call rate and SNP call rate were set at  $\geq 90\%$ . Samples were repeated if  $\geq 10\%$  of SNPs in the plex failed or if  $\geq 50\%$  of SNPs in the plex showed up as aggressive or moderate or no call.

## **3.21. TaqMan® SNP genotyping/allelic discrimination**

### **3.21.1. Basic principle**

TaqMan SNP genotyping is suitable for genotyping a small number of SNPs in a large cohort. The TaqMan SNP genotyping assay amplifies and detects specific SNP targets via two allele-specific fluorescent probes. One probe labeled with VIC dye for the detection of allele 1 and the other probe labeled with FAM dye to detect the other allele. The perfect matched probe is hybridized to the target whereas the mismatched probe is degraded by Polymerase.

By exciting the reporter, the emission of the released reporter can be measured (Figure 3.17). Moreover, a minor groove binder (MGB) probe can increase the differences in melting temperature values between matched and mismatched probes, which allows for more accurate allelic discrimination (Life Technologies).



**Figure 3.17.** Allelic discrimination using TaqMan SNP genotyping probes. (Figure adapted from Perkin Elmer Biosystems)

### 3.21.2. DNA preparation

Two  $\mu\text{l}$  of working DNA [5ng/ $\mu\text{l}$ ] (see section 3.20.2) were transferred onto a 384-well TaqMan plate. As in genotyping with the Sequenom platform, each PCR plate contains one no-template control (NTC) and each assay contains duplicate DNA samples.

### 3.21.3. Primer and probe design

SNP genotype assay mixes for rs11709614 and rs10823418 were pre-designed and commercially provided by Life Technologies. Each assay contained sequence-specific forward and reverse primers to amplify the SNP of interest and two allele-specific TaqMan® MGB probes containing distinct fluorescence to detect specific SNP targets. One was labeled with VIC dye to detect allele 1 and the other was labeled with FAM dye to detect allele 2. The probes used in this study are presented in table 3.8.

**Table 3.8.** TaqMan probes and fluorescent dyes labeled on each allele

SNP ID	Location (NCBI Build 37)	VIC	FAM
rs11709614	Chr.3:24081772	A	G
rs10823418	Chr.10:71439191	C	T

### 3.21.4. PCR step

The reaction needs only three components: 10 ng of genomic DNA, TaqMan genotype master mix, and SNP genotype assay mix, which was prepared with 10  $\mu\text{l}$ /reaction. Because

a robot was used for pipetting, 110 reactions of master mix were prepared for 96 PCR reactions.

Reagent	Volume (µl)	Final concentration
2X TaqMan genotype master mix	2.5	1X
40X TaqMan genotype assay mix	0.125	1X
H <sub>2</sub> O	0.375	
DNA [5 ng/µl]	2	10 ng
Total volume	5	

### PCR cycle

Step	Temperature (°C)	Duration	Number of cycle
Enzyme activation	95	10 min	1X
Denaturation	92	15 sec	40X
Annealing/Extension	60	1 min	

### 3.21.5. Allelic discrimination data analysis

After PCR amplification, an allelic discrimination plate read was performed on the ABI Prism 7900HT Fast Real-Time PCR System (Life Technologies) and analyzed by SDS 2.2.2 software (Life Technologies). Fluorescence intensities were measured during the plate read. The SDS software plotted results of the allelic discrimination based on the fluorescence signals from each well, and showed a scatter plot of alleles 1 versus 2.



## 4. RESULTS

### 4.1. Transcript analysis of the *APC* gene

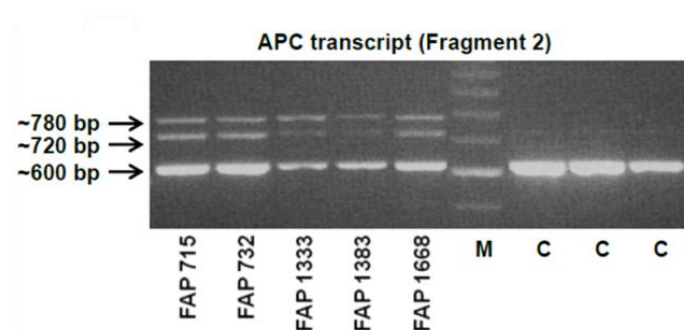
To identify deep intronic *APC* mutations, a systematic *APC* messenger RNA (mRNA) analysis was undertaken in 125 unrelated patients with unexplained adenomatous polyposis meeting the inclusion criteria. The majority of patients were sporadic cases (54%). Clinical details of the patients are summarized in table 4.1.

**Table 4.1.** Baseline data and phenotypic characteristics of the 125 polyposis patients.

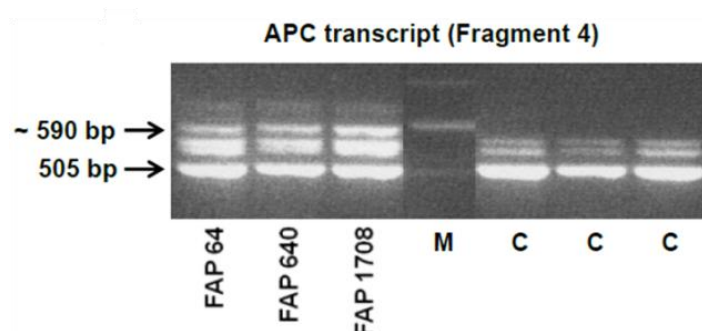
<b>No. of patients</b>	125 patients
<b>Gender (male/female)</b>	75/50
<b>Range of age at diagnosis</b>	20-78 years
<b>Median age at diagnosis</b>	45 years
<b>No. of colorectal adenomas</b>	
< 100 adenomas	66 patients (53%)
> 100 adenomas	27 patients (22%)
Multiple/numerous polyps	32 patients (25%)
<b>Colorectal cancer</b>	39 patients (31%)
<b>Family history</b>	
Familial	20 patients (16%)
Sporadic	68 patients (54%)
Unclear/unknown	37 patients (30%)

#### 4.1.1. Agarose gel electrophoresis

The size and pattern of PCR products from five overlapping fragments spanning the 5'UTR to exon 15A (c.-138 to c.2625) of the *APC* gene were visualized on an agarose gel. Eight of 125 patients (6%) showed an aberrant transcript pattern in one fragment on the gel. The agarose gel of PCR product of fragment 2 clearly revealed two additional bands in five unrelated patients (FAP715, 732, 1333, 1383, and 1668). The bands were approximately 720 bp and 780 bp, exceeding the length of the expected wildtype (~600 bp) (Figure 4.1). The PCR product of fragment 4 revealed an insertion of around 80 bp in three patients (FAP64, 640, 1708) (Figure 4.2).



**Figure 4.1.** Agarose gel representing PCR products of fragment 2 from 5 patients (lanes 1-5) and controls (lanes 7-9). Primers were located in exon 3 (forward) and exon 8 (reverse). All five patients showed two additional bands at around 720 bp and 780 bp compared to controls. M is the DNA marker ladder 100 bp (Invitrogen).



**Figure 4.2.** Agarose gel representing PCR products of fragment 4 from 3 patients (lanes 1-3) and controls (lanes 5-7). The primers were located in exon 10 (forward) and exon 14 (reverse). An additional band around 590 bp was detected in patients but not in controls. M is the DNA marker ladder 100 bp (Invitrogen).

#### 4.1.2. Sanger sequencing of aberrant transcripts and genomic DNAs

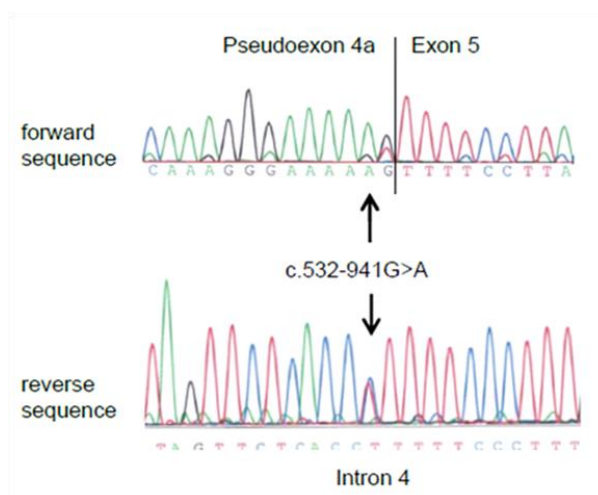
Aberrant patterns on an agarose gel were excised, purified, re-amplified, and sequenced on a 3130xl Genetic Analyzer (Life Technologies). Sequencing of the specific gel bands showed different out-of-frame insertions analyzed by visual inspection with DNA Sequencing Analysis Software version 5.1 (Life Technologies) and by blasting and comparing with the reference sequence to determine the derivation of the insertions.

To identify the causative point mutations which activate intronic cryptic splice sites or create new intronic splice sites, the respective region was subsequently sequenced at the genomic level.

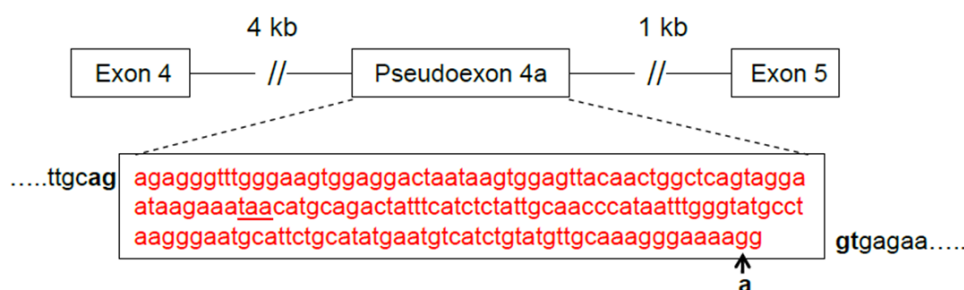
##### Fragment 2 (pseudoexon 4a)

The larger of two additional bands on fragment 2 found in five patients was sequenced. By manual analysis of the sequence, an insertion of 167 bp between exons 4 and 5 of the cDNA (pseudoexon 4a) was identified. By BLAST analysis it was found that the inserted sequence

originated from deep within intron 4 of the *APC* gene. Thus, the description of the mutation on RNA level is r.531\_532ins532-1106\_532-940. A base substitution was visible at the second to last basepair of pseudoexon 4a. Sequencing of the respective region of intron 4 confirmed the heterozygous base substitution (c.532-941G>A) at the second to last base pair of the 3' end of the inserted region (Figure 4.3 and 4.4).



**Figure 4.3.** Sequencing patterns of the largest additional band excised from an agarose gel showing the junction of the 3' end of the insert derived from intron 4 (pseudoexon 4a) and exon 5 (upper panel). Genomic DNA revealed a heterozygous substitution G>A at nucleotide position c.532-941 (reverse sequence) (lower panel).

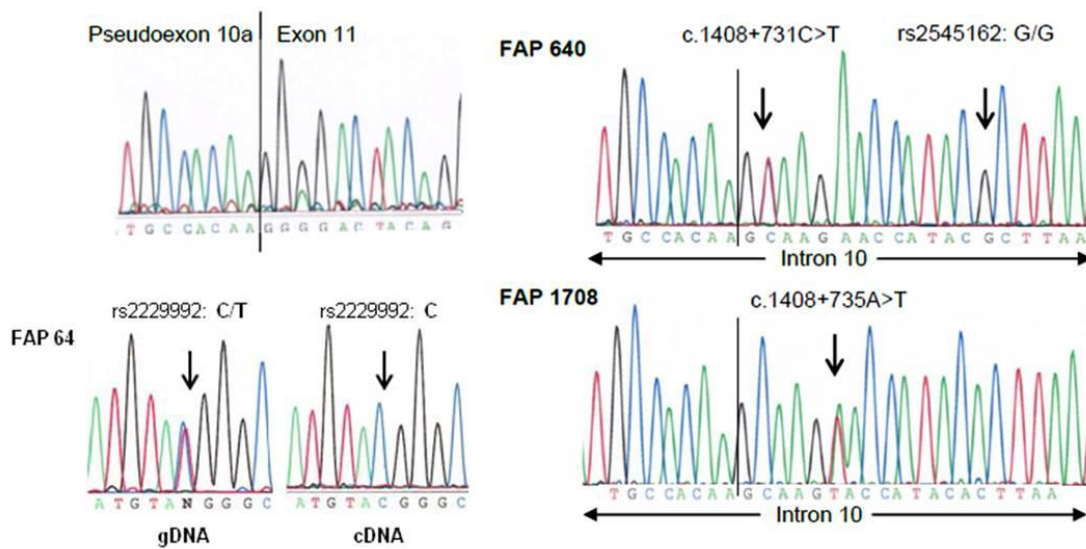


**Figure 4.4.** Diagram representing the pseudoexon 4a deep within intron 4 (top); boxes with numbers denote individual exons. The sequence of the pseudo-exon originating from intron 4 and the flanking intronic sequences with the cryptic splice sites (bold) are shown below. The germline mutation (G>A) is indicated by an arrow, and the predicted premature stop codon (TAA) within the pseudoexon is underlined.

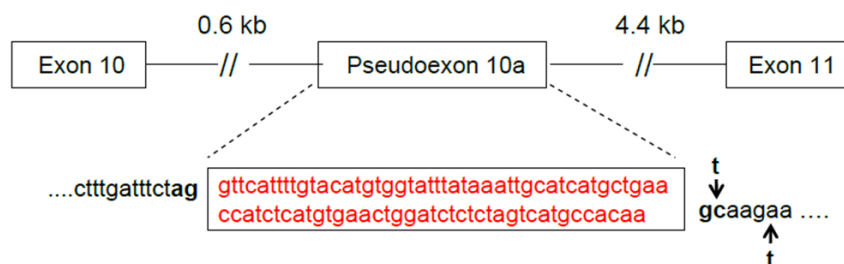
#### Fragment 4 (pseudoexon 10a)

An additional band on fragment 4 from three patients was cut out from the gel, purified, re-amplified, and sequenced. The sequences show the same insertion of 83 bp between exons 10 and 11 of the cDNA (pseudoexon10a), originating from intron 10 of the *APC* gene (r.1408\_1409ins1408+647\_1408+729). The genomic DNA sequence of two patients (FAP64, FAP640) revealed a heterozygous base substitution (c.1408+731C>T) of intron 10 two

nucleotides downstream of the inserted region, which generates a new splice donor site with predicted high splice efficiency. Additionally, the genomic DNA sequence of FAP64 shows a heterozygous genotype for the SNP c.1458T>C; p.Tyr486 (rs2229992) in exon 11 (Figure 4.5). The sequence of the excised wild-type band shows the C allele almost exclusively, indicating a very high splice efficiency of the new splice donor site (Figure 4.6). In addition, patient FAP 64 is heterozygous and patient FAP 640 is homozygous for the known SNP rs2545162 (c.1408+743A>G), located 12 bp downstream of the previously described mutation. Patient FAP1708 presents a heterozygous base substitution (c.1408+735A>T) six nucleotides downstream of the pseudoexon. However, given the length of the insertion, the same splice donor site as in the other two patients was used.



**Figure 4.5. (A)** Sequencing pattern of the largest band (~590 bp) on fragment 4 showing the junction of the 3' end of pseudoexon 10a and exon 11. **(B)** Sequencing of corresponding genomic DNA of intron 10 reveals the heterozygous base substitution c.1408+731C>T together with the homozygous minor allele of SNP rs2545162 (FAP 640) and the heterozygous substitution c.1408+735A>T (FAP 1708). **(C)** Sequencing of FAP64 revealed the heterozygosity of rs2229992 c.1458T>C; p.Tyr486 in exon 11 on the genomic DNA level and the almost exclusive presence of the C allele in the excised wild-type band of fragment 4.



**Figure 4.6.** Diagram representing the pseudoexon 10/11 deep within intron 10 (top) and the sequence of the pseudoexon with the flanking intronic sequences. The cryptic splice sites are shown in bold, and both germline mutations are indicated by arrows.

### 4.1.3. In-silico analysis

The splice prediction program *NNSPLICE 0.9* from the *Berkeley Drosophila Genome Project* (BDGP, [www.fruitfly.org/seq\\_tools/splice.html](http://www.fruitfly.org/seq_tools/splice.html)) was used for calculation of the splicing efficiencies of normal and mutant sequences.

**Pseudoexon 4a:** This occurs in five patients and is flanked by already existing cryptic splice acceptor and donor sites predicted to have high splice efficiency. The substitution increases the splice efficiency of the splice donor site from 83% to 91%. The germline transition c.532-941G>A obviously creates a more canonical splicing signal at the pseudoexon–intron boundary (pseudoexonic part: “AG,” intronic part: “GT”) (Faustino and Cooper 2003) and may thus enhance the efficiency of the splice donor site above a critical threshold. The insertion is predicted to result in a premature stop codon within the inserted region (c.532-941G>A; r.531\_532ins532-1106\_532-940; p.Phe178Argfs\*22).

In comparison to the wildtype sequence, five instead of four A nucleotides were present at the 3' end of the insert indicating an A base substitution or insertion at the second to last position. The variant c.532-941G>A is not recorded in dbSNP/1000Genomes and was not identified in 100 normal controls. Additionally, all five patients presented a heterozygous c.532-845A>G for the known SNP rs77939389 (dbSNP, 1000Genomes), located 94 bp downstream of the 3' end of the insert, with minor allele frequency (MAF) of 0.8-2.0%.

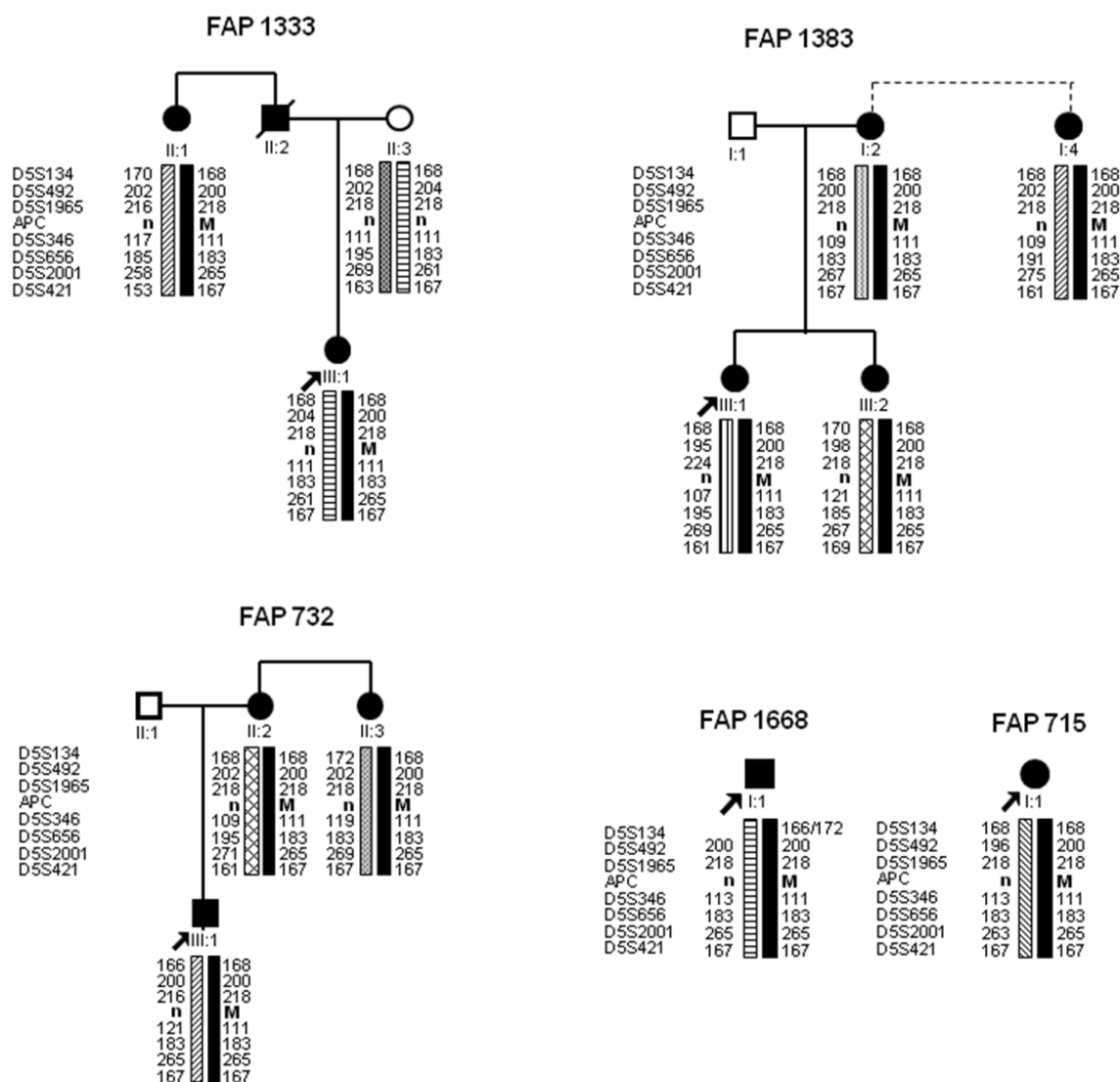
**Pseudoexon 10a:** A heterozygous base substitution adjacent to the 3' end of pseudoexon 10a in patients FAP64 and FAP640 (c.1408+731C>T) generates a new splice donor site with predicted high splice efficiency (99%). The pseudoexon 10a is predicted to result in a frame shift, which leads to a premature stop codon within exon 11 (c.1408+731C>T; r.1408\_1409ins1408+647\_1408+729; p.Gly471Serfs\*55). A putative cryptic splice acceptor site at the 5' end is predicted to have a splice efficiency of 90% (consensus values according to Shapiro and Senapathy (1987)) or 97% (BDGP). The variant is not recorded in dbSNP/1000Genomes and was not found in 100 healthy blood donor controls. For patient FAP1708 who presents a heterozygous base substitution at six nucleotides downstream of the pseudoexon (c.1408+735A>T), in-silico analysis did not point to the generation of a new GT splice donor site secondary to the A>T substitution (splice efficiency 4%), while the CV predicts an increase in splice efficiency from 64% to 69% using the same splice donor region as in the other two patients.

### 4.1.4. Haplotype analysis

All five patients with pseudoexon 4a show a positive family history of polyposis disease. They have an attenuated colorectal phenotype consistent with the established genotype-phenotype prediction in FAP. Haplotype analysis with a panel of seven microsatellite markers flanking the *APC* region was performed in these 5 patients and in seven relatives of three

patients where DNA was available. Testing of healthy relatives showed the wildtype sequence and the major SNP allele c.532-845A in a homozygous state.

According to pedigree information, the five FAP families are unrelated; however, microsatellite analysis demonstrated that all affected members of the families shared the same disease-associated haplotype around the *APC* locus (Figure 4.7), which argues strongly for a founder mutation rather than a mutational hotspot.



**Figure 4.7.** Haplotype analysis in the five families carrying the heterozygous germline mutation c.532-941G>A deep within intron 4 of the *APC* gene using a panel of seven microsatellite markers flanking the *APC* region. All affected persons share the same haplotype marked with a black bar. The haplotypes of index patients FAP 715 and 1668 could not be determined, however, the allele distribution is consistent with the presence of the disease-associated haplotype. The alleles of marker D5S134 in patient FAP 1668 point to a previous recombination event.

## 4.2. CNV analysis

Germline CNV analysis was performed to identify rare, high-penetrant microdeletions and microduplications in novel potentially causative genes. The analysis was carried out in 229 unrelated patients with colorectal adenomatous polyposis (97 females, 132 males) in whom no underlying germline mutation in the known causative genes *APC* and *MUTYH* was identified by routine diagnostics (see section 3.5.1).

### 4.2.1. Quality control of SNP array hybridization

Average overall SNP call rate was calculated by the *GenomeStudio Genotyping* (GT) Module as 99.80% (range 99.19% – 99.90%), and the median was 99.81%. All 229 patients and 531 controls passed quality control with Log R ratio deviation of less than 0.30 (mean 0.232; SD 0.023).

### 4.2.2. CNV calling

QuantisNP calculation was used for calling putative CNVs from two final reports. A total of 296,178 CNVs were called in the control cohort (n = 531), and 103,028 CNVs and 29,473 CNVs were called in the first and second group with 181 and 48 patients, respectively.

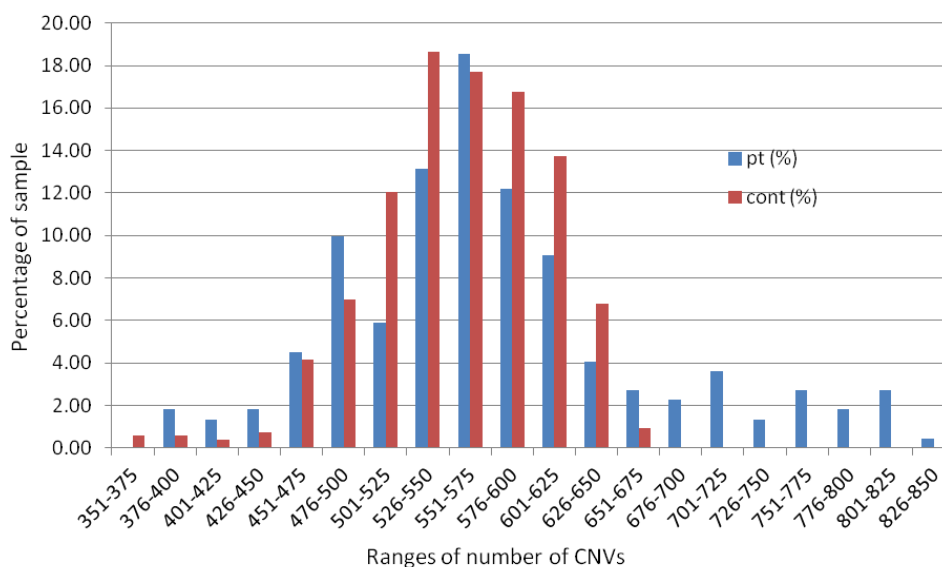
After QuantisNP calculation was performed in all samples, eight patients were excluded from the study since in the meantime a pathogenic deep intronic *APC* mutation was identified by transcript analysis (see section 4.1). Thus, the number of included patients was reduced to 221 (94 females, 127 males) and the total number of called CNVs in the patients was reduced to 127,979 (Table 4.2).

**Table 4.2.** Number of putative CNVs called by the QuantisNP program

Lot	Sample	No. of Sample	Total called CNVs	Mean of CNVs/ sample	SD	Med	Max No. of CNVs	Min No. of CNVs	CNVs smaller than 1 kb
1 <sup>st</sup>	Control	531	296178	558	57.31	559	662	359	79147 (26.7%)
1 <sup>st</sup>	Patient	181	103028	569	94.82	556	855	374	27338 (26.5%)
2 <sup>nd</sup>	Patient	48	29473	614	67.67	584	780	539	9151 (31%)
Eight patients were excluded from the study because deep intronic <i>APC</i> mutations were identified.									
<b>Combined</b>	Patient	221	127979	579	92.76	568	855	374	35299 (27.6%)

The average number of 579 CNVs per patient was not significantly different from that in controls (558 CNVs). The majority of patients (49.5%) and controls (65.2%) carried 501-600 CNVs. Twenty-eight patients (12.6%) carried more than 700 CNVs whereas there was no control who presented more than 700 CNVs (Figure 4.8). The 28 patients with more than 700 CNVs had no obvious specific clinical characteristics compared to patients with less than 700

CNVs. However, the mean number of CNVs  $\geq 10$  kb per individual was the same (109 CNVs) in both patient and control cohorts.



**Figure 4.8.** Distribution of called CNVs in the patient (blue bars) and control (red bars) cohorts; the x-axis presents the ranges of the number of CNVs called per individual and the y-axis shows the corresponding percentage of patients and controls.

The smallest size of a called CNV was 1 bp and the largest size was 3,184 kb; the number of probes varied from 1 to 610 and the max log BF ranged from 0 to 1606.22. The number of CNVs with the same size was consistent between patients and controls. Around 27% of called CNVs were smaller than 1 kb, and defined as ‘indels’. The majority of CNVs (53.5%) in both groups were smaller than 10 kb (Table 4.3). The number of CNVs in each size range is shown in table 4.3.

**Table 4.3.** The percentage of CNVs of each size in patients and controls

Length	Patients (%)	Controls (%)
< 1kb	27.58	26.72
1 - 9.99 kb	53.57	53.73
10 - 49.99 kb	13.73	13.97
50 - 99.99 kb	2.72	2.84
100 - 199.99 kb	1.57	1.79
200 - 499.99 kb	0.63	0.74
500 - 999.99 kb	0.17	0.19
$\geq 1$ Mb	0.02	0.02



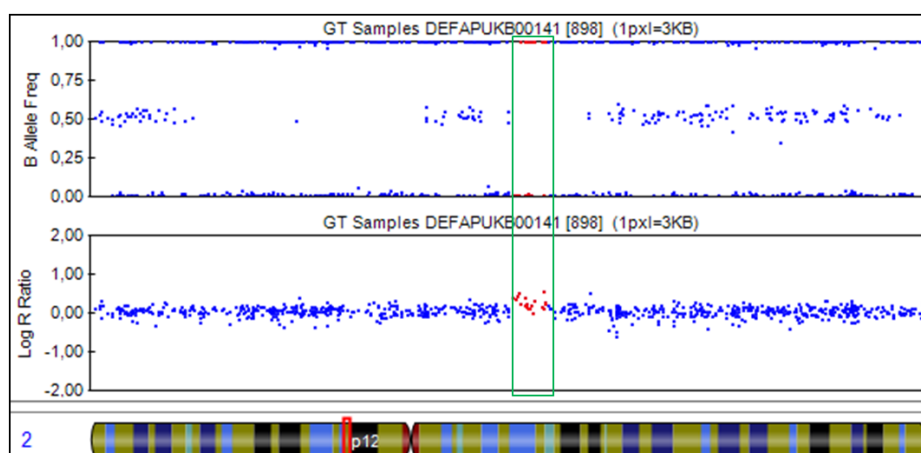
### 4.2.3. Survey of CNV in known candidate genes

Before applying filtering steps, it was determined whether the putative CNVs cover known or established colorectal cancer (CRC) genes (Table 3.4) based on the genomic positions of the genes (NCBI build 36, hg 18). In 44 patients, deletions and/or duplications in the genes *BMPR1A*, *CHD1*, *CTNNB1*, *EGFR*, *MAP2K4*, and *POLE4* were found (Table 4.4).

Two CNVs carrying duplications of *CTNNB1* and *EGFR* passed the filter criteria and were included in the study. Both genes function as oncogenes, thus, overexpression of the gene (gain of function mutation) due to the duplication would be in line with an oncogenic mutation and thus reasonable to be causative.

All other CNVs failed the inclusion criteria because their size or max log BF did not fit the filter criteria, or the CNVs involved only an intronic region. Not only were filter criteria applied; the CNVs were also visually inspected on the Genome Viewer.

The duplication (895-27270 bp) involving part of *MAP2K4* showed a false positive pattern on the Genome Viewer. All seven probes inside the gene were CNV probes and assembled only within 895 bp. For the duplication (71395 bp) encompassing *POLE4*, the Log R Ratio (LRR) was higher than that of the flanking regions and its B allele frequency (BAF) showed loss of heterozygosity (Figure 4.9) instead of a duplication pattern.



**Figure 4.9.** Genome viewer (Illumina GenomeStudio) demonstrating a duplication on chr2:75010111-75081505 with 19 markers, max log BF 15.0471, involving the *POLE4* gene. B allele frequency (BAF) represents no heterozygous probes and Log R Ratio (LRR) shows markers with intensity higher than zero.

**Table 4.4.** Known cancer predisposing genes found in the patient cohort.

Gene	Chro	Postion	No. of pt	Type of CNV	Part of disruption	Size	No. of probe	Max log BF	Remarks
<i>BMPR1A</i>	10q22.3	88506376-88674925	7	Duplication - 5	5'UTR+exon 1	648-1697	12-13	0.5-6.2	excluded because Max logBF failed the inclusion criteria
				Deletion - 2	partial gene	7429	2	4.1-7.1	excluded because Max logBF failed the inclusion criteria
<i>CDH1</i>	16q22.1	67328696-67328696	1	Deletion	intron	538	3	2.7	excluded because Max logBF failed the inclusion criteria, and only an intronic region was involved
<b><i>CTNNB1</i></b>	<b>3p21</b>	<b>41211405-41256943</b>	<b>1</b>	<b>Duplication</b>	<b>whole gene</b>	<b>527160</b>	<b>187</b>	<b>305.322</b>	<b>included in further study</b>
<b><i>EGFR</i></b>	<b>7p12</b>	<b>55054219-55242524</b>	<b>6</b>	<b>Duplication- 1</b>	<b>partial gene</b>	<b>338588</b>	<b>180</b>	<b>440.403</b>	<b>included in further study</b>
				Deletion - 5	intron	193-810	4-12	5-37	excluded because only an intronic region was involved
<i>MAP2K4</i>	17p12	11864866-11987865	28	Duplication	5'UTR+exon 1	895-27270	7-11	0.5-19.6	excluded because Max logBF failed the inclusion criteria
<i>POLE4</i>	2p12	75039283-75050366	1	Duplication	whole gene	71395	19	15.0471	excluded because Max logBF failed the inclusion criteria

#### 4.2.4. CNV filtering

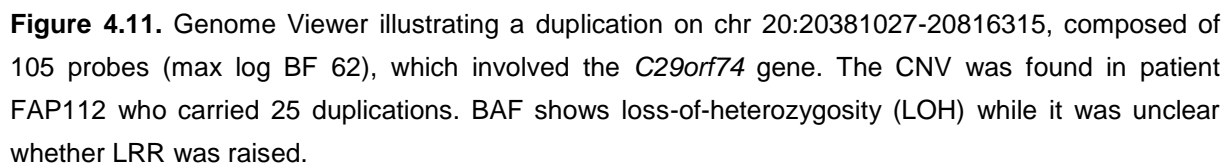
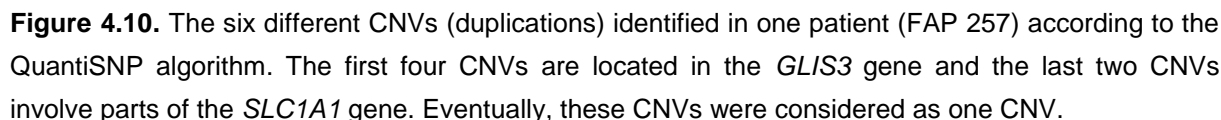
To search for rare, non-polymorphic CNVs, various filter criteria were applied. After filtering by the Cartagenia® software according to size, number of probes, and max logBF, the number of putative CNVs was significantly reduced to 449 deletions and 510 duplications. These 959 CNVs were present in 215 patients. Only 6 patients did not carry any CNVs. The rest carried 1-11 CNVs but there was one patient with 25 CNVs. The majority of patients (60%) carried 2-5 CNVs. The length of CNVs, the number of consecutive probes, the max log BF, and the number of patients with CNVs are summarized in Table 4.5 and 4.6. Of the 959 CNVs, 130 contained no gene. A further 105 CNVs involved only an intronic part of a gene. These CNVs were removed from the study. Therefore, the remaining CNVs consisted of 214 deletions and 461 duplications.

On observation in the *Database of Genomic Variants* (DGV), by visual inspection with the Genome Viewer (Illumina GenomeStudio), and comparison with data from 531 healthy controls genotyped with the same array, it was revealed that 168 of the 214 deletions and 353 of the 461 duplications are common CNVs (found in more than one control) and/or false positive CNVs. These CNVs were removed. Thus, 46 unique heterozygous deletion CNVs and 108 unique duplications (~0.1% of all called CNVs) remained (Table 4.5).

**Table 4.5.** Summary of number of CNVs after each filtering step

Filter criteria	Deletion	Duplication
CNV calling QuantiSNP	89,544	38,435
CNV filtering (Cartagenia)		
≥ 10 kb, ≥ 5consecutive probes, max log BF ≥ 20	449	785
≥ 20 kb, ≥ 7consecutive probes, max log BF ≥ 30	-	510
Further manual CNV filtering		
CNVs without protein coding genes	130	33
CNVs in an intronic region	105	16
CNVs in segmental duplications	0	156
Common CNVs (DGV, HNR cohort) and/or false positive CNVs in Genome Viewer	168	197
CNVs in an outlier patient	0	25
Number of remaining CNVs	46	83

The majority of patients carried only one CNV. However, two patients had considerably more CNVs: patient FAP257 carried 6 duplications located on chromosome 9. The CNVs involve two neighbouring genes separated by few normal markers and thus, they can be considered as one CNV (Figure 4.10). The other is patient FAP112 who carried 25 duplications, all of which showing a loss of heterozygosity (LOH) pattern on the Genome Viewer (Figure 4.11).



**Table 4.6.** Number of CNVs after filtering with the Cartagenia software and after excluding CNVs which failed additional filter criteria (see table 4.5)

87

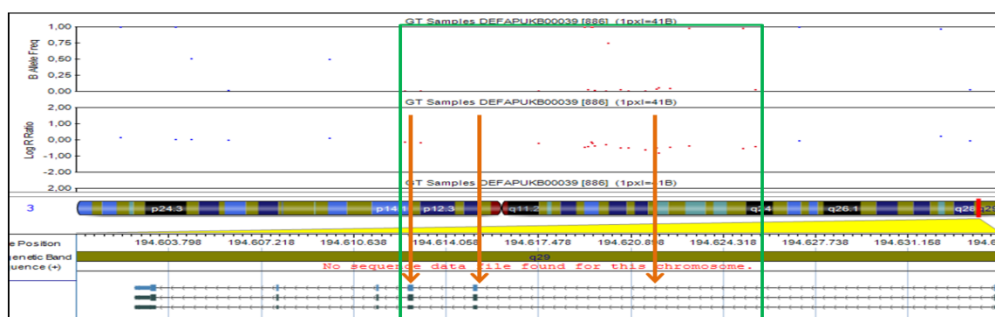
#### 4.2.5. CNV validation by qPCR using SYBR Green I

To validate the copy number of the 46 unique deleted CNVs identified by QuantiSNP, qPCR was applied. 43 of them were validated as deletions. The three unvalidated CNVs showed a normal copy number in exonic regions but a copy number of 1 (deletion) in intronic regions (Figure 4.12). The three false positive deletions were 13-16 kb in size with a max log BF below 40 and involved up to 19 probes. Of the validated 43 deletions, 33 were not present in controls, and 10 were found in one control.

To validate the copy number of the duplications and multiplications, 83 CNVs present in 61 patients were examined by qPCR. 82 of the 83 CNVs were validated and show a gain of copy number. The non validated CNV was located on chromosome X of a female patient. These 82 validated duplications were uniquely present in only one patient; 64 of them were not found in the control cohort, whereas 18 CNVs occurred in one control.

In patient FAP112 with the 25 predicted rare duplications, 4 CNVs were randomly selected for qPCR validation. The Ct calculation showed a normal copy number in all four CNVs. Thus, this patient was excluded.

In summary, a total of 125 confirmed CNVs were present in 93 patients (42%). Each patient carried 1-3 CNVs. Thirty-three CNVs were smaller than 50 kb whereas 62 CNVs were larger than 100 kb (Table A8 and A9).

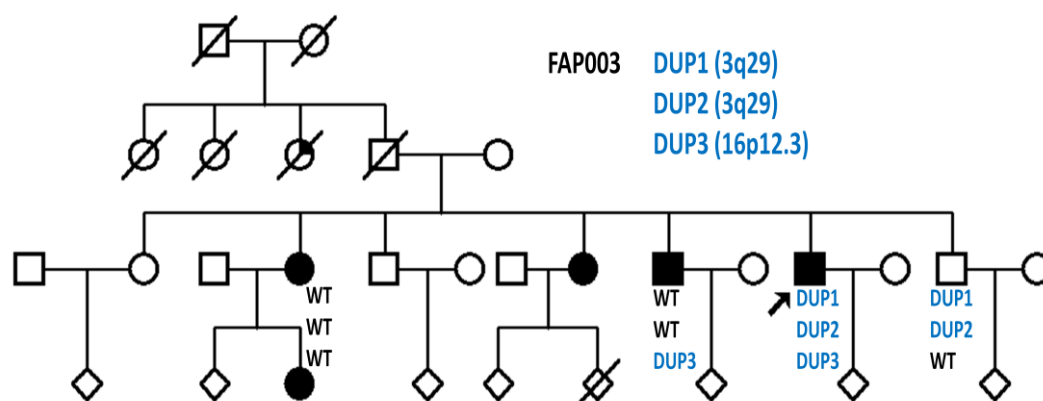


**Figure 4.12.** False positive CNV on chr3:194612501-194625470 (green frame). Three primers were put in two exons and one intron of *ATP13A4* (orange arrows). The  $\Delta\Delta\text{Ct}$  calculation shows normal copy number in the first two probes and loss of copy number in the last probe, which means the two exons are not deleted but the intron is.

#### 4.2.6. Co-segregation analysis

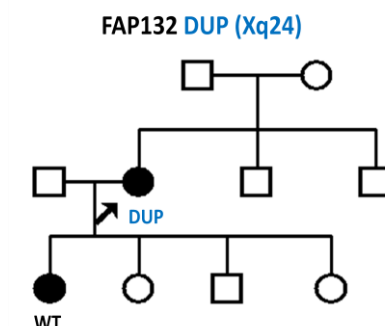
DNA samples of 10 affected and 7 healthy relatives of 5 unrelated index cases were available for examination whether they carry the CNV of the respective index patient. These 5 probands contain a total of 9 CNVs (2 deletions, 7 duplications). Carrier testing was performed using qPCR with the same primers as used for the validation of the CNV in the index patient. From the segregation analysis, we found that none of the CNVs segregated with the phenotype in these five families (Figures 4.13 - 4.17).

## Patient FAP003



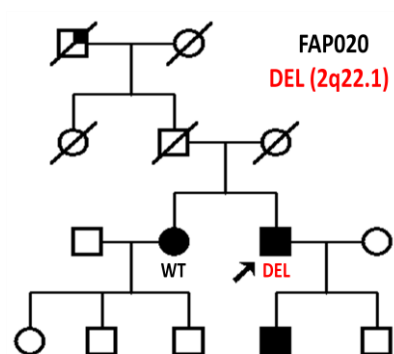
**Figure 4.13.** An affected male carried 3 duplications; two duplications were located on chromosome 3q29 and the third was on chromosome 16p12.3. *C3orf21*, *ACAP2*, *C3orf34*, *PIGX*, *PAK2*, *SENP5*, *NCBP2*, *PIGZ*, *ACSM2*, and *ACSM1* were involved. DNAs of two brothers (one healthy, one affected) and of one affected sister (diagnosed at 38 years old) were used to perform qPCR. The affected sister did not carry any CNV while the affected brother (diagnosed at 30 years old) carried a duplication on chromosome 16 and the healthy brother (44 years old when blood sample was collected) carried two CNVs on chromosome 3. The arrow indicates the index patient (study patient).

## Patient FAP132



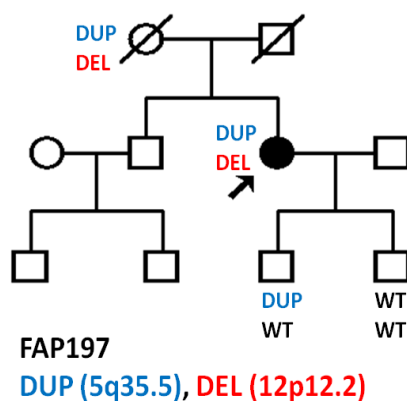
**Figure 4.14.** An affected female carried a duplication on Xq24 (*SLC25A43*, *SLC25A5*). DNA from an affected daughter (diagnosed at 20 years old) was available for copy number analysis. However, she did not carry this duplication. The arrow indicates the index patient (study patient).

## Patient FAP020



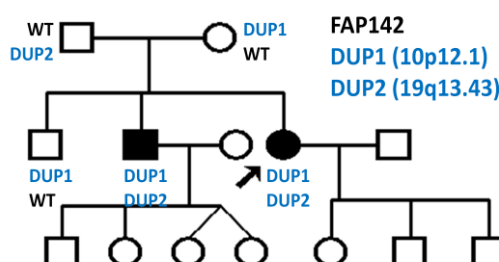
**Figure 4.15.** An affected man carried a deletion on chromosome 2q22.1 (*THSD7B*). A sister of the patient was also affected (diagnosed at 54 years old) but did not carry a deletion on chromosome 2 as in patient FAP020. The arrow indicates the index patient (study patient).

## Patient FAP197



**Figure 4.16.** Patient FAP197 is female and carried two CNVs, a duplication on chromosome 5q35.5 (*BTNL9*) and a deletion on chromosome 12p12.2 (*SLCO1B3*). A mother (unknown age) and healthy sons (16 and 14 years when blood samples were collected) of patient FAP197 were examined whether they carry the same CNVs as the patient. qPCR revealed that the mother carried the same CNVs as the proband and the first son carried a duplication on chromosome 5q35.5. The arrow indicates the index patient (study patient).

## Patient FAP142



**Figure 4.17.** Patient FAP142 is female and carried two duplications, one on chromosome 10p12.1 (*PTCHD3*) and the other on chromosome 19q13.43 (*ZSCAN5A*, *ZNF582*, *ZNF583*, *ZNF667*, *ZNF471*, *ZFP28*, *ZNF470*, *ZNF71*, *ZNF835*). qPCR of the patient's parents and siblings revealed that the patient's affected brother (diagnosed at 58 years old) carried two duplications the same as the patient. However, the unaffected parents (81 and 80 years old when blood samples were collected) and brother (unknown age) also carried one of the two duplications. The arrow indicates the index patient (study patient).

## 4.3. Candidate gene prioritization

### 4.3.1. Genes covered by the validated CNVs

The genes disrupted by the validated CNVs are annotated with RefSeq (hg18) definitions. The 43 deletions involve 68 genes; 66 of them are each deleted in one patient only, but *THSD7B* on chromosome 2q22.1 and *SLCO1B3* on chromosome 12p12 are deleted in two patients. Nevertheless, none of *THSD7B* and *SLCO1B3* CNVs co-segregate with the phenotype in the family of patients FAP020 and FAP197 respectively.

Careful checking with the *Ensembl* browser showed that 33 genes are completely deleted and 35 genes are partly deleted. In *CHL1*, *HSH2D*, *LINGO2*, and *ZNF547*, the deletion involved only an untranslated region (UTR). Twenty genes belong to the Olfactory Receptor (OR) gene family. The 20 OR genes are present in two CNVs, one on chromosome 1q44 and the other on chromosome 12q13.2.

The eighty-two validated duplications involved 168 genes. Fourteen of them failed the inclusion criteria because they are 1) duplicated only in an intronic region (*HBE1*, *HBG2*), 2) duplicated only in an UTR (*KIF26A*, *P2RX1*), 3) located in segmental duplication regions (*ACSM2*, *CCDC74B*, *FAM128B*, *MYH7*, *NUP50*, *POTEF*, *SLC25A5*, *SMPD4*, *TUBA3E*), and 4) found in more than one control (*PPEF1*). Two genes on chromosome 11p15, affected by a whole gene duplication, belong to the OR family. Then 152 duplicated genes remained. 81 of them are partial duplications and 71 genes are whole gene duplications. Each duplicated gene was found in only one patient.



Candidates not fulfilling the inclusion criteria and belonging to the OR family were removed from the study (see discussion). Therefore, the number of candidate genes is reduced as summarized in table 4.7.

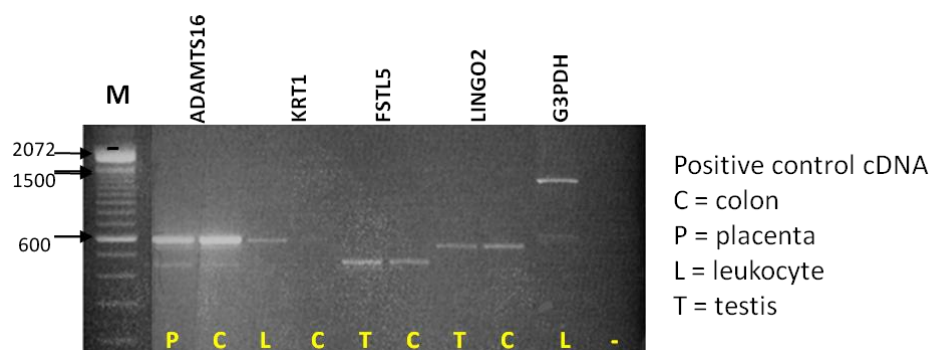
**Table 4.7.** Number of genes involved in candidate CNVs

Results	Deletion	Duplication
No. of involved genes	68 genes	168 genes
No. of excluded genes	24 genes - 4 genes deleted only UTR - 20 genes belong to Olfactory Receptor (OR) family	16 genes - 2 genes located in an intronic region - 2 genes duplicated only UTR - 9 genes located in segmental duplication region - 1 gene found in 7 controls - 2 genes belong to OR family
No. of remaining genes	44 genes - 42 genes presented in 1 pt - 2 genes presented in 2 pts	152 genes (all in 1 pt only) - partial duplication (81 genes) - whole gene duplication (71 genes)

Beside *THSD7B* and *SLCO1B3*, which were found to be deleted in two patients, *DOCK11* on chromosome Xq24 was also present in two patients, one being deleted and the other being partly duplicated. Since *DOCK11* was affected by both a deletion and a duplication, the total number of the remaining candidate genes was 195 genes.

#### 4.3.2. Gene expression in human colon cDNA

According to the expression data on the *EST Profile UNIGENE* database, 129 of 195 candidate genes are expressed in the human intestine. Expression of the 66 genes reported to be unexpressed was examined by PCR with commercial human normal colon cDNA (Ambio). Agarose gel electrophoresis demonstrated that actually 51 of the 66 genes are expressed in colon mucosa by showing a clearly visible band on the gel. 15 of 66 genes did not show a band on the gel although a second primer pair was used for the test and/or a positive control showed expression (Figure 4.18). These 15 genes were removed from the list of 195 candidates.

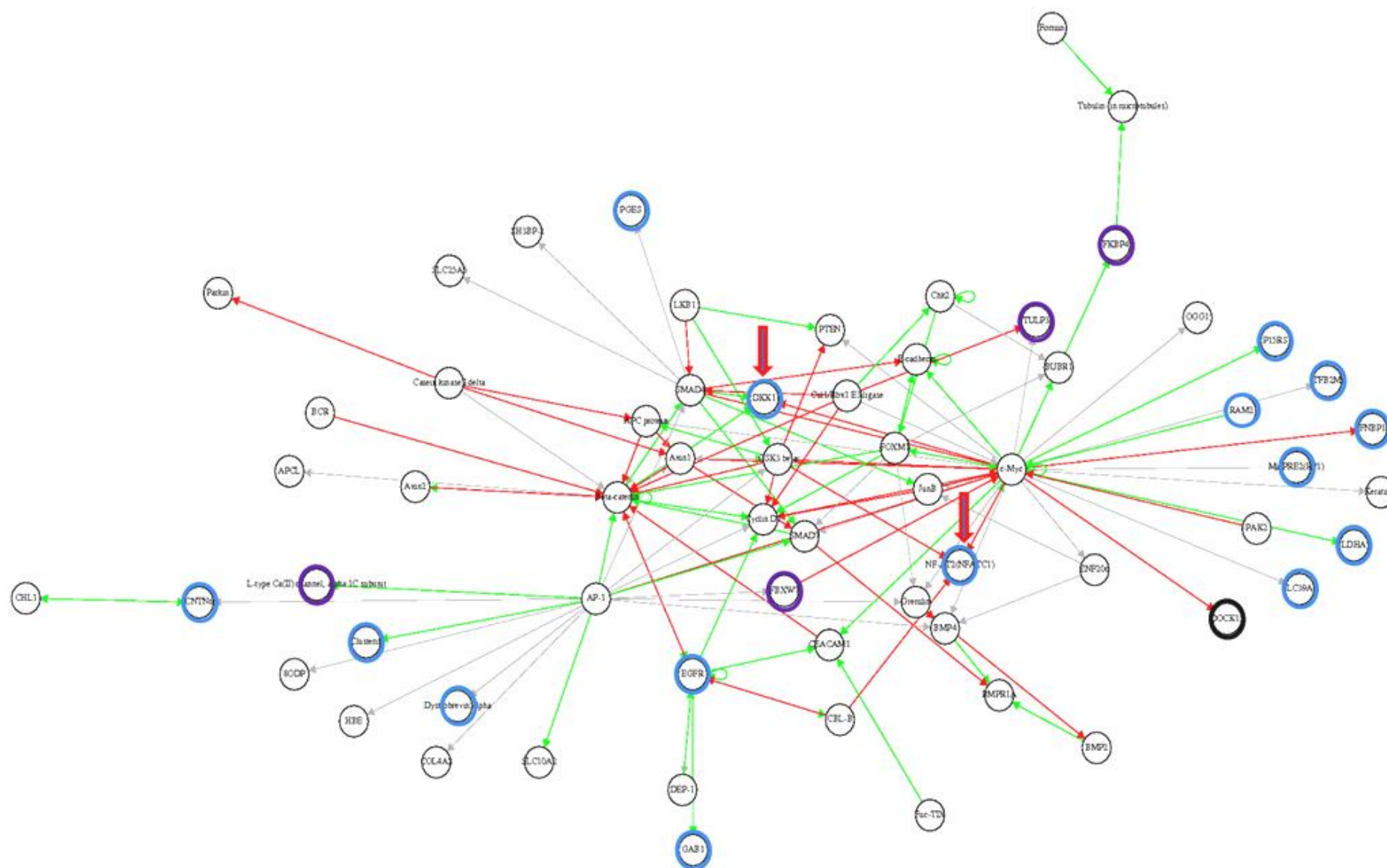


**Figure 4.18.** Agarose gel electrophoresis of candidate genes reported as unexpressed in human intestine (EST Profiles, UNIGENE database). C refers to PCR product of commercial human colon cDNA (Ambio), while P, L, and T are known positive control tissues which express the candidate genes. For example, *KRT1* is expressed in leukocyte cDNA but not in colon cDNA. The last lane is a negative control. GAPDH was used as an internal control to check the efficiency of the PCR reaction.

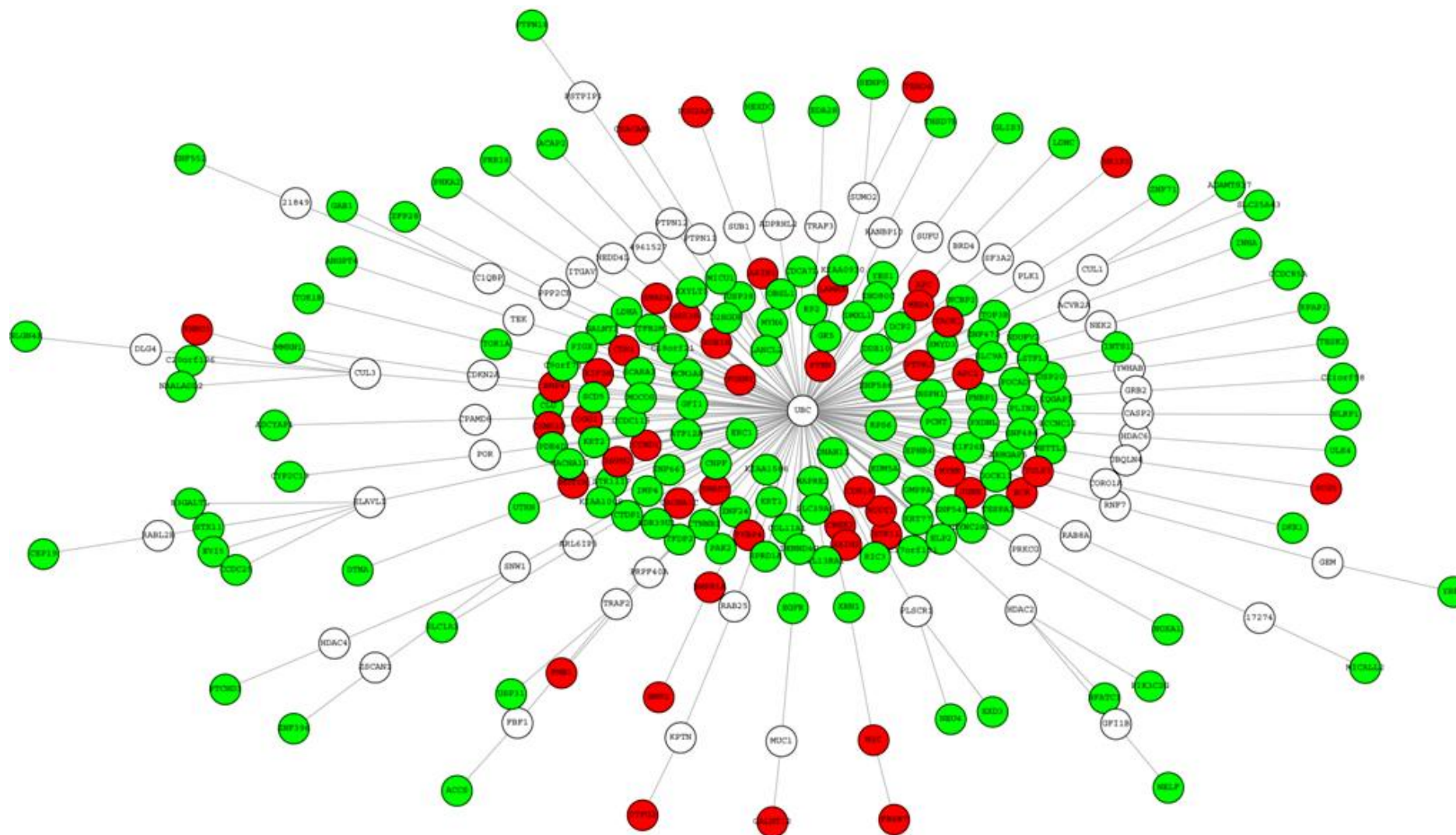
### 4.3.3. Network and pathway analysis

To further prioritize our candidate genes, a network and pathway analysis was performed twice in collaboration with the group of Prof. Fröhlich (Bonn-Aachen International Center for IT (B-IT), Algorithmic Bioinformatics, University of Bonn). For the first analysis, that is, protein-protein interaction, the candidates consisted of 142 genes affected by CNVs of 134 patients and 8 genes from our GWAS analysis (unpublished preliminary results). In addition, 17 established polyposis genes and 26 published candidate genes for CRC were included as anchor or reference genes, respectively. The most interesting candidates are *DKK1* (whole gene duplication) and *NFATC1* (partial gene duplication) (Figure 4.19). Other included CNV candidates which appear in the network are *CDCA7L*, *CLU*, *CNTN6*, *DOCK11*, *DTNA*, *EGFR*, *FBNP1*, *GAB1*, *LDHA*, *MAPRE2*, *PTGES*, *RPRD1A*, *SLC39A6*, and *TFB2M*.

The second network analysis, the 'Steiner tree' algorithm, was performed with all 180 candidate genes. The Steiner tree showed that 135 of the 180 included genes (26 deletions, 61 partial gene duplications and 48 whole gene duplications) are present in the network (Figure 4.20, Table A10). *UBC* was identified in the center of the network.



**Figure 4.19.** Protein-protein interaction analysis. The first analysis of 142 candidate genes from 134 patients (red and blue circles indicate genes affected by heterozygous deletions and by duplications, respectively). Established polyposis genes, published candidate genes for colorectal cancer (black circles), and candidate genes from a polyposis GWAS (purple circles, unpublished data), were used to prioritize putative disease genes. The candidates from the CNV analysis most connected to known genes are *DKK1* (whole gene duplication) and *NFATC1* (partial gene duplication), indicated by red arrows. Thin lines present interaction of the genes: Green color is activating and red color is prohibiting the target gene (pointed by an arrow). Grey edges indicate an unknown mode of action.



**Figure 4.20.** Steiner Tree, the second analysis of 180 candidate genes. *UBC* was identified as a central candidate gene. Green colors represent the candidate genes and red colors indicate putative disease genes.

Subsequently, an enrichment analysis of KEGG pathways and GO terms was performed. The two established oncogenes *CTNNB1*, *EGFR* are involved in many cancer-related pathways. In total, 11 genes were involved in gene sets enriched for cancer related pathways (KEGG database) at  $p < 0.05$  (Table 4.8) and 10 genes were part of a gene set enriched for GO term at  $p < 0.05$  (Table 4.9).

**Table 4.8.** Candidate genes from gene sets enriched for cancer related pathways (KEGG database)

KEGG ID	P-value	Odds Ratio	ExpCount	Count	Size	Term	Candidate gene*
04310	0.000623	5.442095	2.376086	11	150	Wnt signaling pathway	<i>DKK1<sup>w</sup></i> , <i>NFATC1<sup>p</sup></i>
04520	0.002728	7.044397	1.156362	7	73	Adherens junction	<i>IQGAP1<sup>p</sup></i> , <i>YES1<sup>p</sup></i>
04110	0.043709988	3.938282648	1.964230966	7	124	Cell cycle	<i>TFDP2<sup>p</sup></i>
04510	0.04478684	3.134068811	3.168114461	9	200	Focal adhesion	<i>ARHGAP5<sup>p</sup></i> , <i>COL11A1<sup>p</sup></i> , <i>PAK2<sup>w</sup></i> , <i>CCDC148<sup>d</sup></i>

\* *CTNNB1* and *EGFR*, being established oncogenes involved in many cancer related pathways, were not included in the table; <sup>d</sup> deletion; <sup>p</sup> partial duplication;

<sup>w</sup> whole gene duplication

**Table 4.9.** Candidate genes from gene sets enriched for cancer related GO terms (Gene Ontology database)

GOBPID	P-value	Odds Ratio	ExpCount	Count	Size	Term	Candidate gene
GO:2000060	0.002878	90.374603	0.103661	4	7	positive regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process	<i>CLU<sup>p</sup></i>
GO:0051085	0.037736	28.903182	0.148087	3	10	chaperone mediated protein folding requiring cofactor	<i>TOR1A<sup>w</sup>, TOR1B<sup>w</sup>, HSPH1<sup>p</sup></i>
GO:0032880	0.040719	3.176820	4.400097	13	301	regulation of protein localization	<i>ADCYAP1<sup>w</sup></i>
GO:0090090	0.040719	6.520819	1.017164	6	69	negative regulation of canonical Wnt receptor signaling pathway	<i>DKK1<sup>w</sup></i>
GO:0070940	0.040719	134.301887	0.044426	2	3	dephosphorylation of RNA polymerase II C-terminal domain	<i>RPRD1A<sup>w</sup>, RPAP2<sup>d</sup></i>
GO:0032436	0.045831	10.879234	0.427504	4	29	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	<i>CLU<sup>p</sup></i>

\* *CTNNB1* and *EGFR*, being established oncogenes, were not included in the table; <sup>d</sup> deletion; <sup>p</sup> partial duplication; <sup>w</sup> whole gene duplication

### Ingenuity pathway analysis

Another approach for further analysis of the genes obtained from the CNVs analysis was to use the Interaction® pathway analysis tool (Ingenuity Pathway Analysis, IPA). According to the IPA, the 180 candidate genes were involved in 25 independent networks (Table A11). The top four networks consist of 25, 12, 11, and 10 candidate CNV genes (Table 4.10). Top functions of these networks are related to cancer, cellular development, cell death and survival, and cell cycle. The top first network with 25 candidate genes is presented in Figure 4.21.

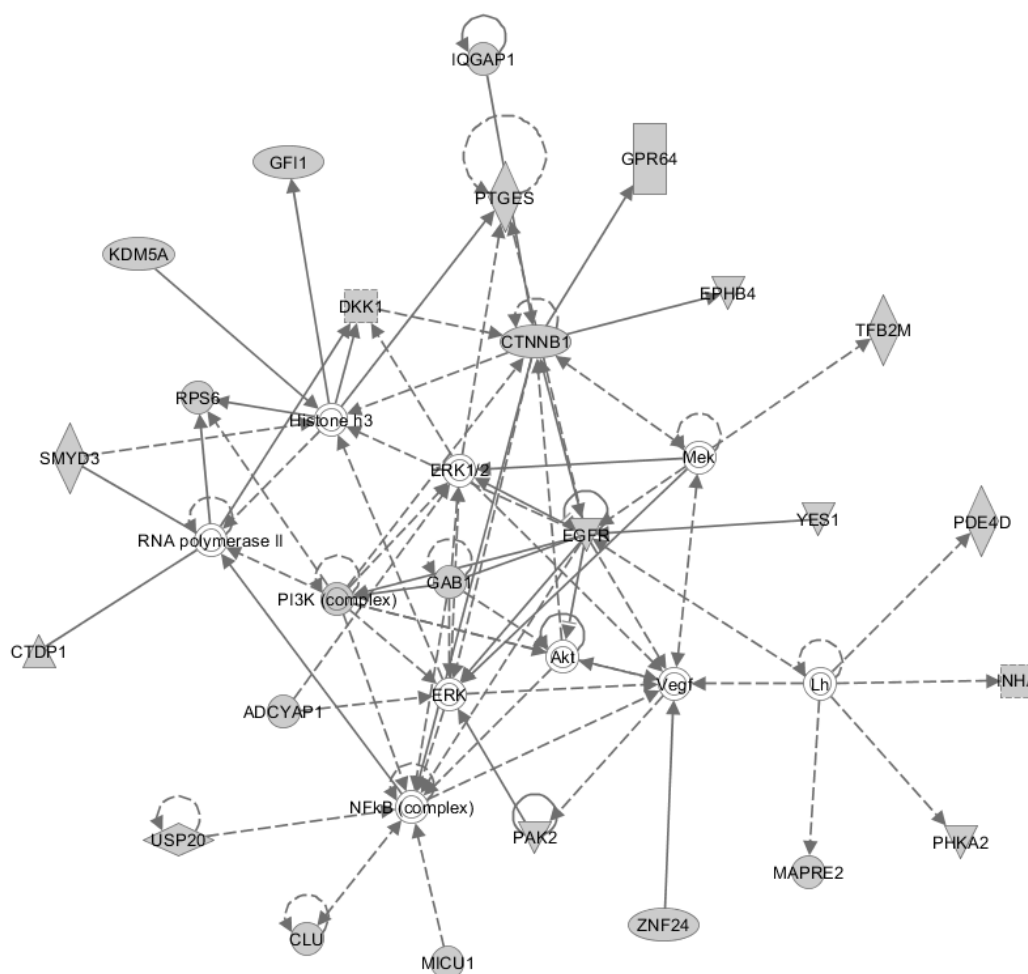
**Table 4.10.** Top four networks of 180 candidate genes and the genes involved in the networks

Net work	Top functions	Genes in network	Score	No. of candidate genes
1	Cancer, Cellular Development, Organismal Injury and Abnormalities	<i>ADCYAP1<sup>w</sup></i> , <i>CLU<sup>p</sup></i> , <i>CTDP1<sup>p</sup></i> , <i>CTNNB1<sup>w</sup></i> , <i>DKK1<sup>w</sup></i> , <i>EGFR<sup>p</sup></i> , <i>EPHB4<sup>p</sup></i> , <i>GAB1<sup>p</sup></i> , <i>GFI1<sup>d</sup></i> , <i>GPR64<sup>w</sup></i> , <i>INHA<sup>w</sup></i> , <i>IQGAP1<sup>p</sup></i> , <i>KDM5A<sup>p</sup></i> ( <i>JARID1A</i> ), <i>MAPRE2<sup>p</sup></i> , <i>MICU1<sup>d</sup></i> ( <i>CBARA1</i> ), <i>PAK2<sup>w</sup></i> , <i>PDE4D<sup>d</sup></i> , <i>PHKA2<sup>w</sup></i> , <i>PTGES<sup>w</sup></i> , <i>RPS6<sup>w</sup></i> , <i>SMYD3<sup>p</sup></i> , <i>TFB2M<sup>p</sup></i> , <i>USP20<sup>w</sup></i> , <i>YES1<sup>p</sup></i> , <i>ZNF24<sup>w</sup></i>	46	25
2	Cellular Development, Hematological System Development and Function, Hematopoiesis	<i>ARHGAP5<sup>p</sup></i> , <i>CMA1<sup>w</sup></i> , <i>COL11A1<sup>p</sup></i> , <i>EVI5<sup>d</sup></i> , <i>HSPH1<sup>p</sup></i> , <i>NFATC1<sup>p</sup></i> , <i>PKD1L2<sup>p</sup></i> , <i>PTGER3<sup>d</sup></i> , <i>SCD5<sup>d</sup></i> , <i>SLC1A1<sup>p</sup></i> , <i>UTRN<sup>p</sup></i> , <i>ZFP28<sup>w</sup></i>	17	12
3	Cell Cycle, Cancer, Cell Death and Survival	<i>ADAMTS17<sup>p</sup></i> , <i>CAMKK1<sup>w</sup></i> , <i>CDCA7L<sup>p</sup></i> , <i>DOCK11<sup>d</sup></i> *, <i>DTNA<sup>p</sup></i> , <i>ERC1<sup>p</sup></i> , <i>LDHA<sup>w</sup></i> , <i>MOCOS<sup>p</sup></i> , <i>NCBP2<sup>w</sup></i> , <i>TFDP2<sup>p</sup></i> , <i>XRN1<sup>p</sup></i>	15	11
4	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry	<i>DGP2<sup>d</sup></i> , <i>IL13RA1<sup>w</sup></i> , <i>MICALL2<sup>w</sup></i> , <i>NLRP1<sup>d</sup></i> , <i>NOXA1<sup>w</sup></i> , <i>PCNT<sup>p</sup></i> , <i>PLIN2<sup>d</sup></i> ( <i>ADFP</i> ), <i>SEN5<sup>w</sup></i> , <i>STX11<sup>w</sup></i> , <i>USP31<sup>p</sup></i>	13	10

\* *DOCK11* is both deleted and partially duplicated; <sup>d</sup> deletion; <sup>p</sup> partial duplication; <sup>w</sup> whole gene duplication

By functional annotations of the 180 candidate genes, IPA showed the two established oncogenes *CTNNB1* and *EGFR*, and in addition, *DKK1*, as the top three genes related to cancer ( $p = 1.67 \times 10^{-5}$ ). Eleven candidate genes, *ARHGAP5*, *CLU*, *CTNNB1*, *DNAH11*, *EGFR*, *EPHB4*, *FSTL5*, *GAL3ST2*, *KIF26B*, *ZNF24*, and *ZNF470*, are related to colon carcinoma and adenocarcinoma ( $p < 0.01$ ) (Table A12).





**Figure 4.21.** Top function network formed by the IPA software, relevant for cancer, cellular development, organismal injury and abnormalities. Of the 46 genes in this network, 25 were candidate genes from our CNV analysis. A simple straight line represents direct interaction, a dash line represents indirect interaction, an arrow head means action upon a gene.

#### 4.3.4. Data mining

To further evaluate the pathogenic relevance of the candidates, various online databases were used for compiling all available information on the 180 candidate genes.

##### Causative monogenic disease

Twenty-seven genes have been reported in OMIM, DAVID, and GENATLAS as being causative for specific non tumor related autosomal dominant or autosomal recessive monogenic conditions (Table A13). These genes were removed from the list for further work up if the specific mode of inheritance, the characteristic symptoms, or the usual age at onset of the hereditary conditions argues against a causative role in colorectal tumor predisposition given the phenotype and family history of our patients.

## GWAS Catalog

Thirty-nine candidate genes (8 deletions, 19 partial duplications, 12 whole gene duplications) have been reported in the GWAS catalog. SNPs in several of them are associated with various cancers but not with CRC, as well as with complex neurodegenerative diseases, immune deficiency diseases, and other traits such as height and smoking behavior. However, since GWAS identify low-penetrant risk alleles, data from the GWAS catalog was not adequate to judge whether the candidate genes should be either included or excluded from the study.

## STRING

Using the STRING database to see protein-protein interactions of our candidate genes with known somatically mutated cancer genes reported in COSMIC at or above medium confidence (0.4), we found that 54 candidate genes are somehow associated with those known cancer genes. Among the candidate genes found to be associated with known CRC genes are *C10orf11*, *EDA2R*, *FSTL5*, *GFI1*, and *TESK2*. All 54 candidate genes and the known somatically mutated cancer genes are presented in Table A14. However, a few of them, including *C10orf11*, cause specific phenotypes and were excluded from the study.

## Publications

For 87 genes an association with cancer has been reported in original studies published in the PUBMED database. In addition to the two known oncogenes (*CTNNB1*, *EGFR*), several genes are notable such as *DKK1*, *EDA2R*, *EPHB4*, *FOCAD*, which are assumed to be candidate tumor suppressor genes. Additionally, several candidates are involved in pathways related to cancer. Particularly, *RSPO4* is involved in the Wnt signaling pathway; *NLRP1*, *TESK2* are involved in apoptosis; and *EDA2R* is involved in the *p53* signaling pathway (Table A15). Twenty-five of them are up-regulated or down-regulated in CRC.

By compiling relevant data, we were able to reduce the number of interesting candidates in several ways; we excluded genes causing other specific but non-tumor-related monogenic disorders; we comprehensively studied genes with functions unrelated to features relevant for tumorigenesis and/or genes expressed at very low levels in human adult colon cDNA; we also studied some genes not segregating with the phenotype in the family. Notably, more than one rule was applied to judge whether a gene should be excluded from the candidate list. On the other hand, we did not exclude genes which do not segregate with the phenotype but are present in more than one patient. The same rule was applied to genes of unknown function such as those belonging to the Zinc Finger Protein family.

After applying these filtering steps, 97 genes derived from the CNV analysis (31 deletions, 45 partial duplications, 22 whole gene duplications, [in *DOCK11* both a deletion and a duplication were identified]) remained (Table A16). In addition, the *UBC* gene, which is centrally located in the network analysis, was included in the priority list. Besides two well-

known oncogenes (*CTNNB1*, *EGFR*), 32 candidates are (i) related to known cancer pathways such as Wnt, ErbB, p53, TGF $\beta$  (*C10orf11*, *CCDC148*, *DKK1*, *EDA2R*, *GAB1*, *NFATC1*, *RSPO4*, *STK11IP*); (ii) reported as a putative TSG (*DKK1*, *EPHB4*, *FOCAD*, *FSTL5*, *LZTFL1*, *PDE4D*, *SCARA3*); (iii) involved in functions relevant for tumor development like cell cycle regulation, proliferation control, apoptosis, adherens junction, cell adhesion (*ARHGAP5*, *CAMKK1*, *CCDC148*, *CNTN6*, *DOCK11*, *EDA2R*, *EVI5*, *GAB1*, *GFI1*, *IQGAP1*, *KIF26B*, *MAPRE2*, *MCM3AP*, *NLRP1*, *PIK3C2G*, *RPS6*, *TESK2*, *TFDP2*, *UBC*, *XRN1*, *YES1*); (iv) related to CRC (*ARHGAP5*, *CLU*, *CTNNB1*, *DNAH11*, *EGFR*, *EPHB4*, *FSTL5*, *GAL3ST2*, *KIF26B*, *ZNF24*, *ZNF470*). *FOCAD* has been identified in early-onset cancer disease, and *TESK2* is located upstream of *MUTYH*.

To further evaluate the plausibility of the remaining 32 candidates, they were examined regarding (i) the occurrence of somatic mutations in colon tumor tissue (COSMIC database), (ii) the likelihood of haploinsufficiency according to the *Haploinsufficiency score* (Huang et al. 2010), and (iii) their genetic intolerance to functional variation according to the *Intolerance score* (Petrovski et al. 2013). High frequencies (> 5%) of somatic mutations in colorectal tumors have been described in 9 of the above 32 genes (*ARHGAP5*, *CNTN6*, *FOCAD*, *FSTL5*, *IQGAP1*, *MCM3AP*, *NLRP1*, *PIK3C2G*, *XRN1*). A low value (< 25th percentile) on the intolerance score, indicating high sensitivity to genetic alterations, has been described in 11 of the 32 genes (*ARHGAP5*, *CAMKK1*, *CNTN6*, *EPHB4*, *IQGAP1*, *KIF26B*, *MCM3AP*, *NFATC1*, *PDE4D*, *SCARA3*, *UBC*). The *haploinsufficiency score*, which indicates dosage-sensitive genes, was very low (< 10%) for three genes (*MAPRE2*, *TFDP2*, *YES1*) and showed moderate sensitivity (< 40%) in another 12 genes (*CAMKK1*, *CNTN6*, *DOCK11*, *EPHB4*, *EVI5*, *GAB1*, *GFI1*, *IQGAP1*, *KIF26B*, *NFATC1*, *TESK2*, *XRN1*). In total, 9 of the 32 genes showed significant scores in two of the three parameters (Table 4.11). *CTNNB1*, *EGFR*, and three genes not mentioned above (*DMXL1*, *KDM5A*, *SLITRK6*) exhibited significant values for all three parameters.

**Table 4.11.** Nine of the 32 candidate genes related to processes relevant for tumorigenesis, which are frequently affected by somatic mutations in colorectal tumors (> 5%), tend to be intolerant to functional variation (Intolerance score < 25<sup>th</sup> percentile), and/or are likely to be haploinsufficient (Haploinsufficiency score < 40%). All are from partial duplications and present in one patient only.

Gene	Chr	Part of gene	Function & pathways/Literatures	% of somatic mutation <sup>a</sup>	Intolerance score <sup>b</sup>	Haploinsufficiency score <sup>c</sup>
<b><i>ARHGAP5</i></b>	14q12	5'UTR + exon 1	<b>Focal adhesion</b>	<b>7.9</b>	<b>20</b>	44
<b><i>CAMKK1</i></b>	17p13.2	whole gene	involved in <b>regulating cell apoptosis</b> , promotes cell survival	1.1	<b>19</b>	<b>13</b>
<b><i>CNTN6</i> *</b>	3p26.3	exon 9-23 + 3'UTR	<b>cell adhesion</b> , Notch signaling pathway	<b>7.4</b>	<b>10</b>	<b>23</b>
<b><i>EPHB4</i></b>	7q22.1	exon 13-17 + 3'UTR	<b>acts as TSG</b> , angiogenesis pathway	3.4	<b>3</b>	<b>20</b>
<b><i>IQGAP1</i></b>	15q26.1	5'UTR + exon 1-2	<b>Adherens junction</b> , interacts with <b>APC</b>	<b>5.9</b>	<b>5</b>	<b>17</b>
<b><i>KIF26B</i> *</b>	1q44	2 exons	<b>positive regulation of cell-cell adhesion</b>	4.5	<b>18</b>	<b>26</b>
<b><i>MCM3AP</i> *</b>	21q22.3	exon 21-28 + 3'UTR	<b>inhibits DNA replication and cell cycle progression</b>	<b>5.5</b>	<b>1</b>	60
<b><i>NFATC1</i></b>	18q23	exon 12-13 + 3'UTR	transcription factor, <b>non-canonical Wnt signaling pathway</b>	3.8	<b>2</b>	<b>38</b>
<b><i>XRN1</i></b>	3q23	exon 30-42 + 3'UTR	<b>involved in homologous recombination, meiosis, telomere maintenance, and microtubule assembly</b>	<b>5.6</b>	31	<b>11</b>

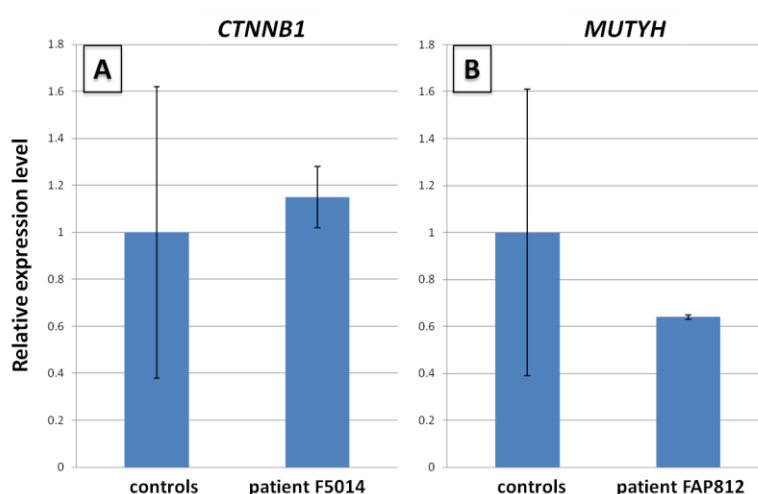
\* in these genes truncating germline mutations have been found in other patients in addition to the duplication identified in the index patient (see section 4.5.2);

<sup>a</sup> COSMIC database; <sup>b</sup> Petrovski et al., 2013; <sup>c</sup> Huang et al., 2010.

#### 4.4. TaqMan® gene expression analysis of *CTNNB1* and *MUTYH*

Patient FAP5014 carried a whole gene duplication on chromosome 3p22.1 involving the canonical Wnt signalling oncogene *CTNNB1*, whereas patient FAP812 harbored a heterozygous deletion on chromosome 1p34, partly covering the *TESK2* gene which is located 3.4 kb upstream of *MUTYH*. To examine whether the expression levels of *CTNNB1* and *MUTYH* were affected by these CNVs, TaqMan gene expression analysis using leukocyte RNA was performed in these two patients comparing them to ten healthy anonymous controls (5 males, 5 females).

The average *CTNNB1* expression level of patient FAP5014 ( $1.15 \pm 0.13$ ) was slightly higher compared to those in the controls ( $1 \pm 0.62$ ) and the average *MUTYH* expression level of patient FAP812 ( $0.64 \pm 0.01$ ) was slightly lower compared to that in controls ( $1 \pm 0.61$ ) (Figure 4.22). Although the mean expression levels of both genes were changed in the expected direction, there was no significant difference between cases and controls ( $p = 0.56$  and  $0.54$ ; Wilcoxon test). Notably, the gene expression levels varied markedly among the controls. In patient FAP812, also no heterozygous *MUTYH* germline point mutation was identified by Sanger sequencing of the coding regions.



**Figure 4.22.** Results of expression analysis of *CTNNB1* in a patient carrying a duplication CNV encompassing the whole *CTNNB1* gene (A) and *MUTYH* in a patient with a deletion CNV partly covering the *TESK2* gene (B). The expression level of the target gene of each patient was compared to 10 anonymous healthy controls (5 males, 5 females). Results are shown as the means of three replicate reactions with standard deviations illustrated as vertical bars. The difference in expression between the control group and the patients was not significant.

## 4.5. Resequencing candidate genes

To further evaluate the pathogenicity of the candidate genes, a mutation analysis was performed to look for germline point mutations in the most interesting ones. We were specifically looking for additional pathogenic germline point mutations in the respective genes assuming a dominant model of inheritance (heterozygous mutations in other patients) and a recessive model of inheritance (biallelic point mutations in other patients or a heterozygous point mutation in the same patient who carried a CNV in the respective gene).

In this study, two methods were used for the mutation analysis. *LZTFL1* was sequenced by Sanger sequencing while the remaining 97 genes were sequenced by high-throughput targeted NGS.

### 4.5.1. Sanger sequencing of *LZTFL1*

The *Leucine zipper transcription factor-like 1* (*LZTFL1*) gene was selected for screening for pathogenic germline point mutations by Sanger sequencing. *LZTFL1* is located on 3p21.3, which is a hotspot for putative TSG and has been reported as a candidate cancer gene (Sjoblom et al. 2006) and as a novel TSG, which may inhibit tumorigenesis by stabilizing E-cadherin-mediated adherens junction formation and promoting epithelial cell differentiation (Wei et al. 2010). Forty-six variants of *LZTFL1* have been reported in the *Exome Variant Server* (EVS) database. The majority of them are not in coding regions. Only one variant, reported in EVS, leads to a truncating mutation of the gene.

The coding regions (10 exons) and the flanking intronic regions of the *LZTFL1* gene were sequenced in 100 randomly selected polyposis patients of the study cohort. The SeqPilot® software (JSI medical systems GmbH, Germany) was used for the analysis of the sequence data. Three variations were detected in 13 patients.

A heterozygous single base substitution located in exon 8, predicted to result in a missense mutation (c.736G>A;p.D246N), was found in eleven patients. The variant was reported in the *dbSNP* database (rs1129183) with a minor allele frequency (MAF) of 6%. By in-silico analysis using *Mutation Taster* and *PolyPhen-2*, the variant was predicted to be benign and possibly damaging, respectively.

Another heterozygous single base substitution located in exon 8 and predicted to result in a missense mutation (c.625A>T;p.S209C) was found in one patient. The variant was reported in the *dbSNP* database (rs148560000) with no MAF data. By in-silico analysis using *Mutation Taster* and *PolyPhen-2*, the variant was predicted to be benign, suggesting a rare polymorphism. The last heterozygous single base substitution, located in the 5'UTR (c.-147C>T), was predicted by *Mutation Taster* to be a polymorphism. The sequencing results are summarized in table 4.12.

**Table 4.12.** *LZFTL1* variants found in 100 polyposis patients and the results of in-silico analysis

No. of pt	DNA change	AA change	Position	dbSNP (MAF)	Mutation Taster-prediction	PolyPhen-2
11	c.736G>A	p.D246N	exon 8	rs1129183 (MAF=6%)	polymorphism	possibly damaging
1	c.625A>T	p.S209C	exon 8	rs148560000 (N/A)	polymorphism	benign
1	c.-147C>T	-	5'UTR	-	polymorphism	-

#### 4.5.2. High throughput resequencing of remaining candidate genes

To validate the etiological relevance of the candidate genes by identifying point mutations in additional patients, the coding exons and adjacent intronic regions of all candidates were sequenced in 100 patients from the primary study cohort and in 92 patients from a second cohort (45 patients with adenomatous polyposis and 47 patients with Lynch syndrome-like phenotype) using targeted NGS (Table 3.1). Based on TruSeq Enrichment assays, all 97 candidate genes, including *UBC*, consisted of 1497 targets and were covered by 2090 probes. The overall performance of the enrichment and sequencing process is shown in table 4.13.

**Table 4.13.** The overall performance of the enrichment and Next Generation Sequencing process \*

Target size	622657
Number of target regions	2163
Total reads	3.98 million (+/- 14%)
Unique mapped reads	3.67 million (+/- 14%)
GC drop out	2.2% (+/- 0.1)
AT drop out	11.6% (+/- 1.5)
Median library insert size	244 (+/- 6)
Read length	2 x 101 (paired end)
Mean coverage	304 (+/- 44)
Covered at depths of 10x	97.7%
Covered at depths of 20x	96.6%
Covered at depths of 30x	95.6%

\* the enrichment assay included a total number of 140 genes, 97 of which are relevant for this study

In total, 356 variants passed the filter criteria. The first analysis focused only on truncating mutations including frameshift mutations, nonsense mutations, and mutations at the canonical splice sites. In-frame and missense variants were not included. The sequence reads of all remaining 26 unique truncating mutations were visually inspected with the

VARBANK Read Browser (Figure 4.23). None of them were removed because the mutated reads were unlikely to be false positives. However, four variants showed a low read depth, only 8 reads each variant, and 2 of 8 reads showed a variation. Five variants were excluded from the study since they were reported in the dbSNP and EVS databases as frequent variants in the general population.



**Figure 4.23.** Snapshot of the VARBANK Read Browser (Cologne Center for Genomics) showing NGS results of exon 13 of the *CNTN6* gene. Red labeled positions (chr 3:1413999-1414000) indicate a heterozygous deletion (roughly 50% of reads are affected) of two bases (CG) which is predicted to result in a frameshift mutation and premature stop codon. Sanger sequencing confirmed the result as a true positive mutation.

The remaining 21 variants were verified by Sanger sequencing. All except the four variants with low read depth were validated as true positive variations. 16 of the 17 true positive variants occurred in a single patient each, except for a frameshift mutation in *PTGER3* (c.1185delC; p.N395fs) that was found in 4 patients. However, the latter variant has frequently been reported in population-based controls (EVS database) with an MAF of 0.87% (A1A1=0/A1R=72/RR=4055). Another variant identified in the *SDR39U1* gene and predicted to result in a frameshift (c.542delG;p.G181fs) is also frequently found in the European general population with an MAF of 0.50% (A1A1=2/A1R=36/RR=3908). Both variants were removed from the list of potential pathogenic germline point mutations (Table 4.14).



**Table 4.14.** Seventeen validated rare truncating mutations found in 192 patients by targeted NGS

Genes	Chro	Positive	Mutated exon (total No. of exon)	Allele1	Allele2	cDNA	Protein	Number of pt.
<i>CNTN6</i>	3	1269501	4 (23)	G	A	c.183-1G>A	CSS	1
<i>CNTN6</i>	3	1363516	8 (23)	A	.	c.944delA	p.Y315Lfs*38	1
<i>CNTN6</i>	3	1413981	13 (23)	A	T	c.1493-2A>T	CSS	1
<i>CNTN6</i>	3	1413999-1414000	13 (23)	CG	.	c.1509_1510delCG	p.V504Pfs*22	1
<i>FOCAD</i>	9	20770046	10 (46)	C	T	c.715C>T	p.Q239*	1
<i>FOCAD</i>	9	20929349	29 (46)	T	A	c.3079-8T>A	CSS	1
<i>HEXDC</i>	17	80400061-80400062	12 (12)	CA	.	c.1353_1354delCA	p.H451Qfs*66	1
<i>HSPH1</i>	13	31719695	11 (18)	C	T	c.1584+5G>A	CSS	1
<i>KIF26B</i>	1	245530424	3 (15)	T	.	c.754delT	p.C252Vfs*86	1
<i>MCM3AP</i>	21	47704082-47704091	1 (28)	GTCAC TTTCT	.	c.1110_1119delAGA AAGTGAC	p.E371Tfs*20	1
<i>PTGER3*</i>	1	71418662	4 (4)	G	.	c.1185delC	p.N395Kfs*9	4
<i>PXDNL</i>	8	52252210	21 (21)	C	A	c.4120G>T	p.E1374*	1
<i>SDR39U1*</i>	14	24909629	6 (6)	C	.	c.542delG	p.G181Afs*22	1
<i>TESK2</i>	1	45812650	8 (10)	C	A	c.792+1G>T	CSS	1
<i>ULK4</i>	3	41439565	35 (37)	C	T	c.3678+5G>A	CSS	1
<i>YBEY</i>	21	47706876-47706877	2 (5)	GC	.	c.51_52delGC	p.P18Tfs*4	1
<i>ZNF471</i>	19	57035963	5 (5)	G	.	c.527delG	p.S176Lfs*18	1

\* frequently found in population-based European controls; CSS, canonical splice site

The remaining 15 heterozygous truncating point mutations (single base substitutions and small deletions or insertions of coding regions or canonical splice sites) were found in 11 genes. No mutation was identified in the *UBC* gene. The 15 rare truncating mutations were present in 15 (8%) of the 192 patients.

In these 11 genes, additional 66 unique missense variants were detected. 53 of 66 variants were found in one patient only while 13 variants were found in 2-31 patients. All 13 variants are frequently reported in the dbSNP and/or the EVS databases and are thus not regarded as high-penetrance mutations. The same was true for 26 of the 53 variants found only once in the patient cohort. The remaining 27 rare missense mutations, each present in a different patient, are listed in Table 4.15. In-silico analysis, using *PolyPhen-2*, *MutationTaster*, and *SIFT* online tools, partly resulted in inconclusive predictions regarding their pathogenic impact. However, 11 variants were predicted to be deleterious by all three tools (Table 4.16).

**Table 4.15.** Summary of putative pathogenic point mutations identified in 11 candidate genes

Gene	Type of CNV	No. of truncating mutations	No. of missense mutations	No. of missense mutations predicted to be deleterious *	Total no. of variants reported in EVS	Truncating mutations reported in EVS
<i>CNTN6</i>	Partial duplication	4	1	0	248	8
<i>FOCAD</i>	Deletion	2	1	0	441	11
<i>HEXDC</i>	Partial duplication	1	2	0	160	4
<i>HSPH1</i>	Partial duplication	1	1	1	161	1
<i>KIF26B</i>	Partial duplication	1	9	5	363	3
<i>MCM3AP</i>	Partial duplication	1	2	2	359	4
<i>PXDNL</i>	Partial duplication	1	4	1	301	10
<i>TESK2</i>	Deletion	1	2	0	119	3
<i>ULK4</i>	Partial duplication	1	3	0	330	10
<i>YBEY</i>	Partial duplication	1	0	0	19	0
<i>ZNF471</i>	Whole gene duplication	1	2	2	85	8
<b>Total</b>		15	27	11		

\* predicted to be deleterious by all 3 in-silico tools (*PolyPhen-2*, *Mutation Taster*, *SIFT*)

In total, 26 rare mutations (15 truncating mutations and 11 missense mutations predicted to be deleterious) in 11 genes were present in 22 patients. One patient had two variants in the same gene (*HSPH1*): however, the phase could not be determined since no DNA was available from relatives. The other patients had only one mutation per gene, and no patient carried both a point mutation and a CNV of the respective gene (Table 4.17).

Among the 11 genes with a deletion or duplication and an additional point mutation, *CNTN6*, *HSPH1*, *KIF26B*, and *MCM3AP* are supposed to be intolerant against genetic variation (low *intolerance score*) and *CNTN6*, *HSPH1*, and *KIF26B* are predicted to be haploinsufficient. Three genes (*CNTN6*, *FOCAD*, *MCM3AP*) are frequently mutated in colorectal tumors (Table 4.11). *CNTN6* has the highest frequency of truncating mutations whereas *KIF26B* exhibits the highest frequency of variants overall and in *MCM3AP* only variants likely to be pathogenic have been identified.

Since oncogenes are often affected by gain-of-function mutations such as specific activating missense mutations or amplifications, we also specifically looked for missense mutations in those genes, which were affected by whole gene duplications and which had been included

in the targeted sequence approach. We identified 23 missense mutations in *CAMKK1*, *CTNNB1*, *DKK1*, *ELP2*, *GAL3ST2*, *KIAA1009*, *PTPN18*, *ZFP28*, *ZNF471*, *ZNF667*, and *ZNF835* (Table A17). Four of the 23 missense mutations were predicted by all 3 in-silico tools (PolyPhen-2, Mutation Taster, SIFT) to be disease causing. Two of the four are found in *ZNF471*, in which a frameshift mutation was found as well. The other two were found in *CTNNB1* and *PTPN18* (Table 4.16).

Focusing on the 32 interesting candidates related to cancers from CNV analysis and the 13 genes in which germline point mutations were identified, we found that seven patients are affected by more than one mutation in these predisposing genes (Table 4.17).

**Table 4.16.** List of 27 missense mutations in the 11 genes where truncating point mutations have been found and 2 missense mutations in genes affected by a whole gene duplication

Gene	Chro	Position	Exon	Allele1	Allele2	cDNA	Protein	Predicted function (PP/MT/SIFT)
<i>CNTN6</i>	3	1424677	18	T	A	c.2218T>A	p.F740I	benign/DC/damaging
<i>FOCAD</i>	9	20923757	24	G	A	c.2951G>A	p.G984E	benign/polymorphism/tolerated
<i>HEXDC</i>	17	80395185	8	C	T	c.845C>T	p.A282V	benign/DC/tolerated
<i>HEXDC</i>	17	80397586	9	C	T	c.979C>T	p.R327C	benign/polymorphism/damaging
<i>HSPH1</i>	13	31724312	8	A	C	c.916T>G	p.F306V	<b>probably damaging/DC/damaging</b>
<i>KIF26B</i>	1	245530364	3	G	A	c.694G>A	p.G232R	benign/DC/damaging
<i>KIF26B</i>	1	245530514	3	G	A	c.844G>A	p.G282R	<b>probably damaging/DC/damaging</b>
<i>KIF26B</i>	1	245847612	11	T	C	c.2336C>T	p.S779L	<b>probably damaging/DC/damaging</b>
<i>KIF26B</i>	1	245849362	12	C	T	c.3077C>T	p.P1026L	<b>probably damaging/DC/damaging</b>
<i>KIF26B</i>	1	245849920	12	G	A	c.3635G>A	p.S1212N	<b>probably damaging/DC/damaging</b>
<i>KIF26B</i>	1	245850019	12	C	T	c.3734C>T	p.T1245M	benign/polymorphism/tolerated
<i>KIF26B</i>	1	245850064	12	C	G	c.3779C>G	p.P1260R	possibly damaging/DC/damaging
<i>KIF26B</i>	1	245861578	13	C	T	c.5995C>T	p.R1999C	probably damaging/DC/torelated
<i>KIF26B</i>	1	245865829	15	G	A	c.6248G>A	p.R2083H	<b>probably damaging/DC/damaging</b>
<i>MCM3AP</i>	21	47662842	25	A	G	c.5300C>T	p.A1767V	<b>probably damaging/DC/damaging</b>
<i>MCM3AP</i>	21	47692529	8	G	A	c.2411C>T	p.A804V	<b>probably damaging/DC/damaging</b>
<i>PXDNL</i>	8	52284471	19	G	A	c.3863C>T	p.P1288L	<b>probably damaging/DC/damaging</b>
<i>PXDNL</i>	8	52321668	17	T	A	c.2516A>T	p.D839V	probably damaging/DC/torelated
<i>PXDNL</i>	8	52321834	17	C	G	c.2350G>C	p.A784P	possibly damaging/polymorphism/damage
<i>PXDNL</i>	8	52336137	14	G	A	c.1793C>T	p.T598M	probably damaging/polymorphism/tolerated
<i>TESK2</i>	1	45811588	10	C	T	c.958G>A	p.E320K	benign/DC/tolerated
<i>TESK2</i>	1	45923396	2	T	C	c.62A>G	p.E21G	benign/polymorphism/damaging

Gene	Chro	Position	Exon	Allele1	Allele2	cDNA	Protein	Predicted function (PP/MT/SIFT)
<i>ULK4</i>	3	41841779	20	G	A	c.1855C>T	p.R619C	probably damaging/DC/torelated
<i>ULK4</i>	3	41925425	17	C	T	c.1597G>A	p.V533I	possibly damaging/polymorphism/tolerated
<i>ULK4</i>	3	41937038	16	G	A	c.1549C>T	p.H517Y	possibly damaging/polymorphism/damaging
<i>ZNF471</i> *	19	57036217	5	C	T	c.781C>T	p.L261F	<b>probably damaging/DC/damaging</b>
<i>ZNF471</i> *	19	57036986	5	G	A	c.1550G>A	p.C517Y	<b>probably damaging/DC/damaging</b>
<i>CTNNB1</i> * <sup>§</sup>	3	41266829	5	T	G	c.500T>G	p.V167G	<b>probably damaging/DC/damaging</b>
<i>PTPN18</i> * <sup>§</sup>	2	131116862	3	A	C	c.259A>C	p.I87L	<b>probably damaging/DC/damaging</b>

\* genes affected by a whole gene duplication in the index patient; PP/MT/SIFT, Polyphen-2/Mutation Taster/SIFT; DC, disease causing; <sup>§</sup> no truncating mutation was identified

**Table 4.17.** Clinical and genetic details of patients with potentially deleterious mutations in at least two of the most promising candidate genes

Patient ID	Gene	Mutation type	Mutation	Exon	Age at Dx	Phenotype	No.of adenomas	Develop CRC	Extracolonic phenotype	Family history
F354	<i>CNTN6</i>	frameshift	c.1509_1510del CG; p.T503fs	13 (23)	27	attenuated	51-100	yes	normal	S/F
	<i>KIF26B</i>	missense	c.3635G>A;p.S1212N	12 (15)						
F1461	<i>KIF26B</i>	frameshift	c.754delT; p.C252fs	3 (15)	61	attenuated	21-50	no	N/A	S
	<i>MCM3AP</i>	missense	c.2411C>T;p.A804V	8 (28)						
F1520	<i>HSPH1</i>	CSS	c.1584+5G>A	11 (18)	52	attenuated	11-20	no	normal	S/F
	<i>HSPH1</i>	missense	c.916T>G;p.F306V	8 (18)						
F1735	<i>PDE4D</i>	CNV	Deletion	exon 1	44	atypical	100-500	no	N/A	S
	<i>NLRP1</i>	CNV	Deletion	5'UTR + exon 1-3						
F1825	<i>MCM3AP</i>	CNV	Partial duplication	exon 21-28 + 3'UTR	46	attenuated	21-50	yes	N/A	S
	<i>YBEY</i>	CNV	Partial duplication	5'UTR + exon 1-3						
F1359	<i>HSPH1</i>	CNV	Partial duplication	5'UTR + exon 1-4	35	attenuated	51-100	yes	normal	S/F
	<i>GFI1</i>	CNV	Deletion	whole gene						
	<i>EVI5</i>	CNV	Deletion	exon 18 + 3'UTR						
F5014	<i>FOCAD</i>	stopgain	c.715C>T;p.Q239*	10 (46)	52	attenuated	51-100	yes	normal	S
	<i>CTNNB1</i>	CNV	Whole gene duplication	whole gene						

CNV: copy number variant; CSS: canonical splice site; S: sporadic case; F: familial; N/A: not applicable

## 4.6. Screening for somatic point mutation

From five patients who carried germline CNVs in candidate TSGs (*DKK1*, *EDA2R*, *EPHB4*, *PDE4D*, and *FSTL5*), Formalin-fixed paraffin embedded (FFPE) polyp tissue was available to screen for somatic “second hits” (point mutations, large deletions) in tumor DNA. The results are summarized in tables 4.18 and 4.19.

**Table 4.18.** Candidate TSGs and results of screening for somatic point mutations in tumor DNA

Patient ID	CNV type	CNV position	Involved gene (TSG)	Part of gene	Sequence result
FAP1576	duplication	chr10:53735827-53798419	<i>DKK1</i>	whole gene	No variation
FAP1427	deletion	chrX:65699679-65934932	<i>EDA2R</i>	whole gene	fail
FAP1245	duplication	chr7:100207480-100244080	<i>EPHB4</i>	exon 13-17+ 3'UTR	variant identified*
FAP1735	deletion	chr5:58578074-58734281	<i>PDE4D</i>	exon 1	variant identified*
FAP764	deletion	chr4:162981340-163113553	<i>FSTL5</i>	exon 4	No variation

\* for details see table 4.19

Two variants identified in *PDE4D* are located in splice sites of exon 9 and 13; one is a known SNP (rs1553114), the other is g.1545822C>T which was predicted by *MutationTaster* to be a polymorphism and by BDGP to affect neither the splice acceptor nor the splice donor site of the respective exon.

Four variants were detected in *EPHB4*, two of them are located in splice sites of exon 6 and were predicted to be polymorphisms and to affect neither the splice donor nor the acceptor site. The other two are located in exon 6 and 14; c.1272T>A;p.P424 is a mutation predicted to be synonymous while c.2372C>T;p.A791V is predicted to be a deleterious missense mutation by all three in-silico tools (*MutationTaster*, Polyphen-2, SIFT).

**Table 4.19.** Somatic variants identified in the candidate TSGs

Gene	Position	Variation	Protein	Mutation Taster	PolyPhen-2	SIFT	BDGP
<i>PDE4D</i>	g.1531323G>A		SS	Polymorphism*	n/a	n/a	-
<i>PDE4D</i>	g.1545822C>T		SS	polymorphism	n/a	n/a	No change
<i>EPHB4</i>	g.7579C>T		SS	polymorphism	n/a	n/a	No change
<i>EPHB4</i>	g.7918T>A	c.1272T>A	p.P424	disease causing	n/a	tolerated	No change
<i>EPHB4</i>	g.7958C>T		SS	polymorphism	n/a	n/a	No change
<i>EPHB4</i>	g.20968C>T	c.2372C>T	p.A791V	disease causing	damaging	damaging	-

\* rs1553114; AA, amino acid; n/a, not applicable; SS, splice site

## 4.7. Replication of the GWAS performed in adenomatous polyposis

To confirm new potential susceptibility loci identified in mutation-negative adenomatous polyposis patients by a genome-wide association study (GWAS) (unpublished data, not shown here), a replication study of the most promising 119 SNPs (top down approach using the most significantly associated SNPs) was performed in a large Dutch cohort of healthy controls and mutation-negative polyposis patients presenting with the same phenotype as the patients used for the original GWAS. 950 DNA samples (380 cases, 570 controls) were provided by our collaborator in the Netherlands. We excluded five patients (DNA provided in duplicate) and 24 controls (DNA likely contaminated). Therefore, 921 samples (375 cases, 546 controls) remained for the genotyping.

The Sequenom platform was used to genotype the 119 SNPs in the 921 samples. The assay design was able to include all but one SNP (rs10823418) in four plexes. However, one of the 118 SNPs (rs11709614) failed genotyping on the Sequenom platform. Therefore, rs11709614 and rs10823418 were genotyped with the TaqMan SNP genotype assay.

Of the 119 SNPs, 117 SNPs showed a call rate higher than 90% (mean call rate 98.4%) whereas rs1529017 (on plex 3) and rs10922106 (on plex 4) showed a mean call rate lower than 90%. In the four plexes, 11, 5, 10, and 5 samples failed genotyping, respectively. Of 921 samples, 4 samples completely failed in all four plexes, 1 sample failed in 2 plexes, and 13 samples each failed in one of the four plexes. Thus, 903 samples were successfully genotyped for all four plexes.

A preliminary analysis was performed using logistic regression association analysis with call rate of 99% in which rs1529017 and rs10922106 were excluded. Two of the 117 SNPs (rs4236978 and rs797517) showed nominal significance ( $p = 0.004$ , and  $p = 0.008$ ,) with an odd ratio of 1.29 and 1.38, respectively. However, after Bonferroni correction none of the SNPs showed a significant association with the phenotype anymore (Table A18). A comprehensive statistical analysis of the replication data will be performed in collaboration with the IMBIE at the University Hospital Bonn but has not yet been completed.



## 5. DISCUSSION

### 5.1. Transcript analysis of the *APC* gene

In recent years, pathogenic deep intronic point mutations leading to subsequent pseudoexon activation have been identified in a number of different genes and phenotypes including several hereditary tumor syndromes (Beroud et al. 2004; Clendenning et al. 2011; De Klein et al. 1998; Dehainault et al. 2007; Homolova et al. ; King et al. 2002; Nichols et al. 2005; Tuffery-Giraud et al. 2003; Tuohy et al. ; Zhang et al. 2008) and other monogenic conditions such as Becker muscular dystrophy, cystic fibrosis, or hemophilia (Bagnall et al. 1999; Beroud et al. 2004; Chillon et al. 1995; Highsmith et al. 1994; Tuffery-Giraud et al. 2003). These observations indicate that deep intronic splice mutations are more relevant than previously thought; however, to our knowledge this mutation type has so far not been described in FAP patients.

In this study, we performed a transcript analysis in a large number of well characterized patients with clinically proven colorectal adenomatous polyposis, in whom no germline mutation in the *APC* or *MUTYH* genes were detected by routine procedures. By examination of reverse transcriptase (RT)-PCR products of five overlapping fragments spanning coding exons 1-15A of the *APC* gene, we identified a reproducible and intensive aberrant transcript pattern in 8 (6%) of the 125 cases pointing to an underlying intronic mutation on the genomic level. Sequencing of the aberrant bands revealed transcript insertions between two exons originating from exonized sequences (pseudoexons, cryptic exons) deep within the corresponding intron. All insertions are predicted to result in out-of-frame transcripts with subsequent premature stop codons, which strongly argue in favor of pathogenicity.

**Pseudoexon 4a:** A 167 bp insertion originating from intron 4 was found in five apparently unrelated patients. The insert is flanked by pre-existing putative cryptic splice donor and acceptor sites with predicted high splice efficiency. The heterozygous germline transition c.532-941G>A at the penultimate position of the inserted sequence creates a canonical splicing signal at the pseudoexon-intron (Faustino and Cooper 2003) and thus might enhance the efficiency of the splice donor site over a critical threshold. The attenuated colorectal phenotype of the patients, which is in line with the established genotype phenotype prediction in FAP (Friedl and Lamberti 2001; Newton et al. 2011), and the segregation analyses in three families further suggest that the mutation is disease causing. Microsatellite analysis demonstrated that the alteration is a founder mutation, which originated on a haplotype shared by all affected persons in the families, rather than a mutational hotspot.

**Pseudoexon 10a:** A genomic transversion c.1408+735A>T at the less-conserved +6 position of the cryptic splice donor site (GCAAGA to GCAAGT) obviously activates the splice site as well as a substitution c.1408+731C>T at the highly conserved +2 splice site position. The donor matrix of the GC–AG group (GCAAGT) shows a remarkable higher degree of conservation than compared with the canonical donor matrix of the GT–AG group (GTAAGT) at positions +3 to +6 (Burset et al. 2000). The sequence motif GCAAGA in *APC* intron 10 can be changed into a splice donor site with high splice efficiency by either a C>T transition at position +2 or an A>T transversion at position +6. The almost 100% splice efficiency of the activated splice donor site was demonstrated by sequencing the normal transcript, where only the wildtype sequence is detectable. The same mechanism has been described in genes underlying other hereditary tumor syndromes such as the *MSH2* gene causing Lynch syndrome (Clendenning et al. 2011) and the *RB1* gene causing familial retinoblastoma (Dehainault et al. 2007).

Nontruncating single-base substitutions in the coding *APC* sequence or unique variants in less conserved intronic regions close to the splice sites have rarely been reported in FAP. Functional studies at the mRNA level have indicated that most of these *APC* variants are pathogenic due to aberrant splicing (Aretz et al. 2004; Kaufmann et al. 2009). In-silico analysis suggests that intronic sequences contain numerous putative splice donor and acceptor sites, which can result in pathogenic pseudoexon activation due to upstream or downstream mutations creating complementary splice sites (Clendenning et al. 2011). Regarding the observed transcript pattern and the segregation within families, these germline mutations are most likely disease causing. Our study points to the existence of a mutational hotspot region in intron 10 and a founder mutation in intron 4 of the *APC* gene. It will be interesting to see whether or not these mutations can also be identified in polyposis patients from other populations.

## 5.2. Novel causative gene identification

### 5.2.1. CNV analysis

In patients with adenomatous polyposis, conventional methods including Sanger sequencing and MLPA are widely used for the detection of germline mutations in the *APC* and *MUTYH* genes. However, in up to 50% of the patients, *APC* and *MUTYH* mutations could not be identified by these methods. Nevertheless the presence of dozens or more colorectal adenomas argues in favour of an underlying genetic predisposition. There might be a few causes of unexplained polyposis due to unrecognized mutations in the established genes that may not be identified by routine molecular methods. The presence of deep intronic *APC* splice mutations (see previous section) emphasize that routine diagnostic tests are not sufficient to detect all causative mutations. Alternatively, mutations in other genes might be relevant, for example, in genes involved in the Wnt signaling pathway. Following the

discovery of *MUTYH* in 2002, it is reasonable to assume that there may still be un-identified causative genes awaiting discovery. Due to the clinical appearance and the pedigree pattern of mutation-negative patients, both monogenic subtypes and a more complex genetic etiology (multifactorial, polygenic cause) seem to be reasonable.

Genomic structural variants, particularly copy number variants (CNVs), are thought to play an important role in human phenotypic variation. CNVs such as germline deletions and duplications are associated with inherited genetic disorders including familial cancer (Lucito et al. 2007). Several studies have explored the frequency of CNVs in well known high-penetrant cancer predisposing genes, including MMR genes, *APC*, *BRCA1*, *BRCA2*, and *VHL* genes. These studies revealed that CNVs, in particular heterozygous deletions of single exons up to whole genes, occur in about 4–15% of the families, which has turned DNA dosage analysis such as MLPA into an essential component of mutation screening in patients with cancer predisposition syndromes (Aretz et al. 2007b; van Hattem et al. 2008; Venkatachalam et al. 2008).

CNVs can predispose to disease by directly affecting the gene dosage of haploinsufficient genes, by unmasking a recessive mutation on the other allele, by changing gene expression due to the disturbance of regulatory regions or inducing epigenetic changes or position effects, or they may function in combination with other genetic and environmental factors (Feuk et al. 2006). Similar to SNPs, the more common CNVs may have no impact on cancer risk or act as low-penetrant risk factors (Shlien and Malkin 2010) while rare deletions and duplications may act as high-penetrant mutations. During the last years, several studies identified both common recurrent and rare heterogeneous germline CNVs as presumptive risk factors for sporadic solid malignancies (Diskin et al. 2009; Liu et al. 2009; Stadler et al. 2012). In addition, it was shown that rare CNVs are collectively overrepresented in familial tumor syndromes compared to controls, and thus contribute to the yet missing heritability (Krepischi et al. 2012a; Pylkas et al. 2012; Shlien et al. 2008).

Therefore, it is reasonable to assume that CNVs, particularly heterozygous deletions might be part of the mutation spectrum in yet unidentified genes responsible for colorectal adenomatous polyposis syndromes. In a mutation-negative FAP family, Thean et al. (2010) identified in all examined polyps a copy number loss at 3q26.1, which might regulate the expression of an upstream candidate TSG (*PPM1L*). However, the authors failed to identify a causative germline variant. De Voer et al. (2013) found a heterozygous microdeletion of *BUB1* in a cohort of early-onset CRC, and subsequently additional point mutations in *BUB1* and *BUB3*.

In the last few years, several calling algorithms have been developed to identify CNVs at the whole genome scale using high-resolution single nucleotide polymorphism (SNP) chips used to perform GWAS. SNP arrays have successfully been employed to identify CNVs in several phenotypes including familial CRC (Yang et al. 2013). These studies clearly demonstrate the

potential of high-resolution genomic copy number profiling for the discovery of new disease genes.

Based on these notions, we performed a high resolution genome-wide SNP-based CNV analysis using Illumina Human Omni1-Quad Bead-arrays to search for novel predisposing genes in 221 well characterized unrelated adenomatous polyposis patients (94 females, 127 males) in whom no mutations in the *APC* and *MUTYH* genes could be determined by routine diagnosis. The majority of patients are sporadic cases without extracolonic lesions and less than 100 colorectal adenomas. The clinical and family features of the patients are consistent with published data of other mutation-negative cohorts (Hes et al. 2014; Mongin et al. 2012; Thirlwell et al. 2007).

We identified altogether 127979 CNVs (579 CNVs per patient). The average number of CNVs was similar in patients and controls. The majority of them carry 500-600 CNVs. Around 27% of these called CNVs are smaller than 1 kb, which are defined as 'indels' and around 53% of the CNVs are smaller than 10 kb, which is in accordance with the size distribution of CNVs in the Database of Genomic Variants (DGV).

Assuming a dominant or recessive mode of inheritance with high penetrance mutations as a monogenic etiology, the frequency of causative CNVs in the general population would be expected to be much lower than 1%. We used stringent filter criteria regarding size, number of probes, and max log BF, to reduce false positive results and to identify rare non-polymorphic CNVs, which are not present in a large control cohort (531 population-based samples). By this approach, the proportion of called CNVs was reduced to less than 0.1% of total called CNVs. CNV deletions were less frequent than duplications (449 deletions, 785 duplications). The number of duplications was further reduced to 510 CNVs when we used more stringent filter criteria. This is in contrast to case-control studies, where chromosomal deletions were more prevalent than duplications (Buizer-Voskamp et al. 2011; Dauber et al. 2011). The number of CNVs was dramatically reduced after applying further filtering steps such as exclusion of CNVs in segmental duplication regions, exclusion of CNVs not encompassing coding genes, and exclusion of common CNVs reported in the DGV and/or found in more than one HNR control. Finally, 46 rare germline deletions and 108 rare germline duplications remained. We eventually validated 43 deletions and 82 duplications as true positives in 42% (93/221) of the patients. All these CNVs were non-recurrent and affected protein coding genes.

CNV calling is a crucial step in CNV detection with genotyping arrays. PennCNV and QuantiSNP are the most popular CNV calling algorithms originally developed for the Illumina platform. Dellinger et al. (2010) reported that QuantiSNP outperforms six other algorithms (Circular Binary Segmentation (CBS), CNVFinder, CNVPartition, Gain and Loss Analysis of DNA (GLAD), Nexus algorithms, and PennCNV). However Xu et al. (2011) compared the calls generated by PennCNV (Wang et al. 2007) and QuantiSNP (Colella et al. 2007) and found that, for the Illumina Infinium 1Mduo chip, CNVs called by PennCNV are generally

shorter and more frequent than those called by QuantiSNP. Later on, Marenne et al. (2012) compared results from the Illumina Human 1M chip processed with *CNV partition*, *PennCNV*, and *QuantiSNP*. They reported that QuantiSNP and PennCNV provided a similar mean number of copy number changes that was higher than that provided by CNVPartition.

Based on these studies and our own experience from previous CNV analyses (Herms et al. 2013) we used QuantiSNP as the only algorithm in this study. Both QuantiSNP and PennCNV are based on the same HMM algorithm, showed a striking overlap of called CNVs and are known to have the best performance in the target size range of CNVs relevant for this study. An alternative approach would have included multiple algorithms to select either only those CNVs which were called by all tools or all CNVs detected by any of the programmes. The former would have resulted in a high rate of true CNVs being missed and the latter procedure in a high rate of false positives (Jiang et al. 2013).

Based on the observation that the validation rate of deletions is higher than that of duplications when the same filtering criteria are applied (Shaikh et al. 2009), we further adjusted the filtering criteria for duplications according to suggestions made by Venkatachalam et al. (2011) to avoid a high number of false positive CNVs. Consistent with Itsara et al. (2009), we found that the majority of deletions (401 of 449 CNVs = 89%) are smaller than 100 kb and more frequent than duplications, of which most are larger than 100 kb (363 of 510 CNVs = 71%). Itsara et al. suggested that the relative enrichment of deletions smaller than 100 kb may reflect higher *de novo* occurrence rates of deletions. In contrast, the rather low numbers of large deletions may be due to their more deleterious effects compared with those of duplications of the same size.

The vast majority of our patients harboured one CNV only, which is similar to published findings in familial and early-onset colorectal cancer (Venkatachalam et al. 2011) and breast cancer (Krepischi et al. 2012a). All CNVs in our study were smaller than 900 kb, consistent with the absence of intellectual disability or other syndromic features in this cohort.

Many studies report CNVs of genes belonging to the olfactory receptor (OR) family. The OR genes are affected at a high frequency by stop mutations in the human population. However, these potential loss-of-function mutations obviously do not have any phenotypic consequences or cause clinical symptoms (Petrovski et al. 2013). Graubert et al. (2007) reported that CNVs of OR genes are involved in environmental response but are not associated with a disease phenotype. Additionally, Cooper et al. (2007) found that CNVs associated with segmental duplications are highly enriched in genes involved in sensory perception including olfactory receptors and components of the immune response. Duplicated regions can appear on the same or different chromosomes (Lander et al. 2001). Therefore, it is not possible to design specific primers to confirm CNVs within duplicated regions. Itsara et al. (2009) excluded segmental duplications from the analysis to avoid ascertainment bias in the results.

Taken together, OR genes seem to be tolerant against point mutations yet liable to a high degree of structural variability. Based on studies mentioned above and the associated methodical problems, we excluded CNVs spanning OR genes and also CNVs located in segmental duplication regions.

In this study, one patient presented with an outlying number of CNV duplications. We randomly examined four CNVs by qPCR but we were unable to verify those CNVs. Together with the CNV pattern shown in the Genome Viewer, we concluded that these CNVs are actually copy-neutral LOH (see section 2.4.2) and thus, the patient was excluded.

By qPCR we verified true positive CNVs in 43 of 46 deleted CNVs and 82 of 83 duplicated CNVs. These numbers indicate that the false positive rate of CNVs called by QuantiSNP is low once they passed our stringent filter criteria and visual inspection. The visual check, taking into account the pattern of LRR and BAF and the type of the consecutive probes of each CNV, could reduce the number of false positive CNVs significantly. The three deletions that could not be validated showed a normal copy number in the coding regions but loss of copy number in intronic parts. The invalidated duplication was on chromosome X of a female. This false calling might be caused by a disadvantage of hybridization-based technologies which include reliance on an accurate reference assembly, difficulty in precise delineation of CNV breakpoints due to noise inherent in measurement of fluorescence (Itsara et al. 2009).

We performed a segregation analyses in 5 patients who carried 9 CNVs in total. However, none of the 9 CNVs co-segregated with the phenotype in the family. These CNVs might either not be causative or contribute to disease manifestation just as low or moderate penetrance risk factors. This is in line with genome-wide CNV studies in other familial tumor syndromes. Krepischi et al (2012b) even suggested that the vast majority of CNVs have moderate penetrance and contribute modestly to human diseases.

From 125 rare germline CNVs, there are six overlapping CNVs in 3 regions (2 CNVs in each region). Two of the six CNVs do not co-segregate with the phenotype. Apart from the 3 pairs of overlapping CNVs, our work is consistent with the very few systematic genomewide CNV studies in unexplained hereditary tumor syndromes, which found as a characteristic feature almost only nonrecurrent CNVs and very little overlap of affected genes. Lucito et al. (2007), Venkatachalam et al. (2011), and Krepischi et al. (2012a) performed genome-wide CNV analysis in unrelated patients with pancreatic cancer, early-onset CRC, and early-onset breast cancer, respectively. In the study of familial pancreatic cancer, several disrupted genes formed a *TP53* centered network and exhibited functions related to genomic integrity (Lucito et al. 2007). In other studies, the filtered genes were not functionally related. All rare CNVs are non-recurrent, and the results of segregation analysis are often inconclusive or argue against a high penetrant CNV contribution. An exception is a study of Yang et al. (2012) in a melanoma-prone family where a duplication, containing interesting candidate genes, segregates with the phenotype in three affected individuals. Recently, Yang et al (2013) found a deletion CNV on 12p12.3 in two of 384 familial CRC patients. Because of the

limited recurrence data obtained so far, it is impossible to state explicitly which of the variants contribute to cancer predisposition. Nevertheless, four of the genes or gene groups affected by CNVs in our study cohort (*ARHGAP5*, *FOCAD*, *KIF26B*, *YBEY*) have also been identified in the above mentioned CNV studies, which supports their pathogenic relevance for cancer predisposition.

These and our non-recurrent results are in accordance with the observation that mutations in newly identified genes underlying inherited tumor predisposition syndromes are very rare. For example, germline mutations in *RAD51C* that cause hereditary breast and ovarian cancer have been found in less than 0.5% of 1,100 breast cancer families or in 1.3% of those 480 families which presented with breast and ovarian cancer (Meindl et al. 2010), and mutations in the polymerase genes, *POLD1* and *POLE*, have been observed in 0.3% of 3,800 families with multiple colorectal adenomas and carcinomas (Palles et al. 2013). At least some of the unexplained tumor syndromes seem to be genetically extremely heterogeneous and large patient cohorts are needed to validate candidate genes by recurrent germline mutations.

### 5.2.2. Candidate gene prioritization

The 125 rare germline CNVs affect 68 deleted genes and 168 duplicated genes. By applying further rational filtering criteria at the gene level, we reduced the number of candidates using the same exclusion criteria for CNVs as for genes, that is, 1) occurring in intronic regions only; 2) located in segmental duplications; 3) belonging to the OR family; and 4) found in more than one control. Subsequently, we excluded genes, which are unexpressed in colonic mucosa. Thereby, the number of candidates was significantly reduced from 236 to 180 genes.

It is quite obvious that the deletion of a gene leads to a loss-of-function even though not all loss-of-function mutations are deleterious, some of them are even advantageous (Conrad et al. 2010). Therefore, as a general rule, the loss of genomic material is more likely to be pathogenic rather than the gain of genomic material as the genome is thought to be more tolerant to duplications than deletions (Brewer et al. 1999). Regarding duplications, it is not unequivocally clear whether they lead to a gain-of-function (e.g. due to overexpression) or a reduction of the expression of the gene.

Whole gene duplications are usually thought to act as gain of function mutations in the sense of an increased expression, and those mutations are unlikely to exist as germline mutations due to serious consequences. However, even if an overexpression can be confirmed, the functional relevance is difficult to validate. In contrast, partial gene duplications may result in deleterious loss-of-function effects depending on their genomic orientation and localization. A loss-of-function effect of a partial duplication was reported e.g. for the *PTPRJ* gene by Kuiper et al. (2010). They found that an intragenic duplication of *PTPRJ* encompassing the

transcriptional control elements and exons 1 to 11 resulted in allele-specific epigenetic silencing of the wild-type *PTPRJ* gene.

In this CNV study, we found on chromosome 1p34 a heterozygous deletion of *TESK2* which is located ~3.4 kb upstream of *MUTYH*. The chromosomal region 1p34 is complex and rapidly evolving (Makalowska 2008). The *TOE1* gene is located between, and shows partial overlap with, *MUTYH* and *TESK2*. Recent work has suggested that an equine SNP in exon 4 of *TOE1* may down-regulate *MUTYH* expression by affecting a transcription factor binding site (Brault et al. 2011). Long-range effects of CNVs on gene transcription due to disruption of regulatory regions or epigenetic modifications are well known (Ionita-Laza et al. 2009; Stranger et al. 2007). Kuiper et al. (2010) identified a disease-causing heterozygous deletion at the 3' region of *EPCAM* located upstream of *MSH2* gene in patients with Lynch syndrome and unexplained *MSH2* loss in tumor tissue. This deletion resulted in a transcriptional read-through of the *EPCAM* transcript and mono-allelic silencing of *MSH2* due to *cis* hypermethylation of the *MSH2* promoter. Jaeger et al. (Jaeger et al. 2012) identified a duplication spanning the 3' end of the *SCG5* gene and a region upstream of the *GREM1* which causes ectopic *GREM1* expression, resulting in a mixed polyposis phenotype.

Therefore, we hypothesized that the *TESK2* deletion might decrease the *MUTYH* expression. We performed an expression analysis using RNA of the affected study patient. However, although the expression level of *MUTYH* in the patient with the *TESK2* deletion was lower than that in controls, the difference was not statistically significant. Albeit CNVs may have a long-range regulatory effect on cancer gene transcription up to 1.2 Mb away from a gene (Stranger et al. 2007), this study could not confirm an effect of the upstream deletion on *MUTYH* expression. Moreover, no heterozygous *MUTYH* point mutation was identified in this patient, thus it is unlikely that in this case the CNV is the underlying cause of polyposis. Nevertheless, it cannot be ruled out that the *TESK2* deletion might have an independent effect on the phenotype itself.

Two well known oncogenes were identified in our study, *CTNNB1* and *EFGR*. *CTNNB1* ( $\beta$ -catenin) is an adherens junction protein and a central player of the canonical Wnt signaling pathway, in which it activates target genes for cell proliferation by interaction with transcription factors. As reported in the *Cancer Gene Census* (The Cancer Genome Project: CGP), it is involved in many types of cancer including adenomatous polyposis (IPA results). It is one of the most common initially altered genes in sporadic colorectal tumors (Pendas-Franco et al. 2008). Consistent with the oncogenic properties of the protein, the whole gene duplication and a missense mutation found in our study would be in line with a gain-of-function alteration leading to overexpression. However, gene expression analysis in the patient who carried the whole gene duplication of *CTNNB1* gene showed no significant increase of the *CTNNB1* expression. This is in consistent with the report of Stranger et al (2007) that 10% of all known duplication CNVs from the International HapMap project show a negative correlation to transcript expression. Furthermore, this patient also harbored a



mutation in *FOCAD*, a novel TSG (see section 5.2.3). Thus, this CNV is either unlikely to be the underlying cause or a high ranking candidate gene for adenomatous polyposis.

The *EGFR* (epidermal growth factor receptor) signaling pathway is commonly activated in CRC, and *EGFR*-target therapies have improved the outcome for CRC patients. The copy number of *EGFR* is increased in human colorectal carcinogenesis and an upregulation of *EGFR* may correlate with malignant progression (Flora et al. 2012). In this study we identified in one patient a duplication, which partly involves *EGFR*. Assuming that the partial duplication disrupts the gene and thus has a loss of function effect, the expression of *EGFR* identified in our study should be decreased, which would not correspond to known oncogenic properties. On the other hand, if the duplication disrupts a silencer, the expression might be increased, and thus the result would be in line with the expected pathological effect. We did not further study the expression pattern of *EGFR* in the patient who carried the partial duplication since *EGFR* seems to be involved in tumor progression rather than initiation of tumorigenesis (Huang et al. 2013; Neumann et al. 2013; Zuo et al. 2013).

The first run of a network analysis, performed by our collaborating colleagues, has shown two interesting candidates associated with known polyposis genes, *DKK1* and *NFATC1*. *DKK1* is also present in the IPA as one of the top three genes associated with cancer, next to *CTNNB1* and *EGFR*. *DKK1*, or dickkopf homolog 1, on chromosome 10q11.2, acts as a tumor suppressor in human CRC cells harbouring endogenous mutations in the Wnt/beta-catenin pathway (Pendas-Franco et al. 2008; Sato et al. 2007), so that loss of *DKK1* may facilitate tumorigenesis. We found a whole gene duplication of *DKK1* which is not in line with the expected inactivating (loss-of-function) mutations typical for TSGs. *NFATC1*, located on chromosome 18q23, is a transcription factor that regulates T-cell development, osteoclastogenesis, and macrophage function. Oikawa et al. (2013) reported that an expression of *NFATC1* contributes to tumor progression due to an increase of invasive activity. This is likely to be a gain-of-function effect. However, in our study, we found a partial duplication of *NFATC1*. Assuming that a partial duplication rather leads to a loss than a gain of function, then the mutation in our patient is not consistent with previous study. Nevertheless, according to previous reports, *NFATC1* is likely to be involved in tumor progression, rather than tumor initiation.

To better understand if these genes are up- or down-regulated by a partial duplication, expression analyses could be performed. Nevertheless, since there were no recurrent CNVs and no germline truncating point mutations of *NFATC1* and *DKK1* were identified in the validation cohort, we did not perform functional studies.

The second network analysis indicated that *UBC*, or Ubiquitin C is located in the center of the network (Steiner tree analysis). Ubiquitination is associated with protein degradation, DNA repair, cell cycle regulation, kinase modification, endocytosis, and regulation of other cell signaling pathways. Its function depends on Lysine (Lys) residue conjugation. *UBC* is involved in the Peroxisome proliferator-activated receptors (PPARs) signaling pathway and is

expressed in the human intestine. Since ubiquitination is a very common and widespread mechanism, it might not be surprising that *UBC* appears in the center of this analysis as the lowest common denominator associated with the highest number of candidates. Together with the absence of a germline mutation in any of our patients, the gene was disregarded from further work-up.

To summarize the network analysis, 3 notable candidates (*DKK1*, *NFATC1*, *UBC*) were identified in two analyses. However, after including all sources of information including patterns of expression and NGS results (see section 5.2.3), these three genes do not seem to be among the most interesting candidate genes for colorectal adenomatous polyposis.

The data mining step was very helpful to reduce the number of candidate genes considerably from 180 to 98. Primarily, we excluded the candidate genes which cause other monogenic, early-onset diseases. If the type of mutation (e.g. truncating) in our patients was consistent with those reported in these disorders, then the study patients should be affected by these phenotypes in addition to the colorectal polyposis. Secondly, several genes were not included in our top candidate gene list because they are well known genes which are obviously not associated with tumorigenesis according to their functions and the pathways they are involved. Instead, data mining could help to select those genes as most interesting candidates whose functions and pathways have been reported to be related to cancer. In addition, we included a group of genes, most of whom belong to the zinc finger protein family, which have unknown functions and properties since genes of which nothing is known about cannot be excluded as causative genes in general.

Literature review of genes predisposing to CRC using an exome sequencing approach demonstrates no overlap among new candidate genes in each study although the same phenotype was studied. Gylfe et al. (2013) identified rare germline truncating mutations in 11 novel susceptibility genes (*UACA*, *SFXN4*, *TWSG1*, *PSPH*, *NUDT7*, *ZNF490*, *PRSS37*, *CCDC18*, *PRADC1*, *MRPL3*, and *AKR1C4*) in familial CRC cases. They are absent or rare ( $MAF \leq 0.001$ ) in the general population. Smith et al. (2013) identified 5 potential novel TSG which predispose to CRC; *FANCM*, *LAMB4*, *LAMC3*, *PTCHD3*, and *TREX2*, by seeking second hit (somatic) mutations in patients who carry germline truncating mutations. None of these novel candidates were found in our study.

The data mining and pathway analyses revealed that 32 candidate genes are of special interest since they are related to well known cancer pathways, are discussed as potential TSG, or are involved in functions relevant for tumor development such as cell cycle regulation, proliferation control, apoptosis, or cell adhesion. Our finding is in line with a study of Fanciulli et al. (2010) which reported that germline copy number changes can affect several gene pathways, including the ERBB2, epidermal growth factor receptor (EGFR) and PI3K pathways in CRC.

Nonetheless, there are other possibilities to further evaluate clinical relevance such as the haploinsufficiency theory, the intolerance score, and frequency of somatic mutation in colorectal tumors.

Assessment of haploinsufficiency (HI) by an HI score is an approach to identify genes, which are likely to be damaged by heterozygous loss-of-function mutations of one allele; i.e. the expression of one allele only (around 50% of the gene product) is not sufficient to maintain the physiological (normal) function (haploinsufficiency). The likelihood of haploinsufficiency of a gene helps to estimate the functional relevance of heterozygous loss of function mutations (deletion CNVs, truncating point mutations). Huang et al. (2010) reported that the most obvious pathogenic mechanism for heterozygous loss-of-function mutations (such as large rare deletions) is haploinsufficiency. Thus, many genes implicated in dominant diseases are supposed to be haploinsufficient. Several disorders are caused by heterozygous germline deletions at several haploinsufficient gene loci such as intellectual disability caused by haploinsufficiency of *ARID1B* (Hoyer et al. 2012). The tendency to delete one rather than both copies might be because critical genes nearby cannot be homozygously deleted (Greenman 2012). Chayka et al (2009) studied the role of clusterin in tumor development by using mouse models of neuroblastoma and they found that clusterin is a haploinsufficient TSG in neuroblastoma. For a haploinsufficient gene functioning as TSG, the loss of even one copy can initiate tumorigenesis instead of the more common haplosufficient TSG, in which a second mutational hit on the wildtype allele is needed to impair gene function (Lambertz et al. 2010; Vasanthakumar et al. 2013).

Petrovski et al. (2013) found that genes responsible for Mendelian diseases are significantly less tolerant to functional genetic variation than genes not causing any known disease. For example, According to the intolerance score the *APC* gene (score 0.90) is intolerant against genetic variation while the tolerance for heterozygous mutations of the haplosufficient gene *MUTYH*, which causes a recessive disease is high (score 62.1). This is in line with their respective patterns of dominant and recessive inheritance models. The intolerance ranking system is helpful for ranking the candidate genes.

Somatic mutations identified in tumor tissue are listed in the COSMIC databases. The higher the frequency of (specific) somatic mutations in a gene which are not present in the general population, the larger is the likelihood for them to be driver genes and relevant to tumorigenesis.

By combining these three parameters (haploinsufficiency score, intolerance score, and number of somatic mutations) and all gathered data, we found that *ARHGAP5*, *CNTN6*, *EPHB4*, *IQGAP1*, *KIF26B*, *MCM3AP*, *NFATC1*, and *XRN1* are the most interesting candidates in the group of 98 genes as they are related to tumorigenesis, involved in cancer pathways, likely to be haploinsufficient genes, frequently affected by somatic mutations, and they tend to be intolerant to functional variations.

Nevertheless, since all approaches mentioned above could not be sufficient to prove the causality of the candidate genes, mutation analysis was performed in parallel to look for recurrent mutations in patients to strengthen the causal relevance of the predisposing genes.

### 5.2.3. Validation of the clinical relevance of the candidate genes

To validate the clinical relevance of the most promising 97 candidate genes, which are either genes of unknown function or known genes with features related to CRC or other cancer types, we screened a large validation cohort (n = 192) for germline point mutations using an NGS-based targeted sequencing approach. Beforehand, the *LZTFL1* gene was screened by Sanger sequencing in 100 polyposis patients.

We identified 15 unique rare truncating mutations of 11 genes in 15 patients and were able to validate all of them by Sanger sequencing. In two genes, multiple mutations were found: *CNTN6* showed a different truncating mutation in each of four patients and *FOCAD* showed a different mutation in two cases. All other truncating mutations were found only once. Subsequently, 11 missense mutations predicted to be deleterious were identified in these 11 genes. Besides one patient with a putative compound-heterozygous *HSPH1* mutation, no case with a biallelic point mutation or a heterozygous point mutation unmasked by a CNV was found, indicating that autosomal recessive inheritance is not frequent in genes affected by rare CNVs.

Our strategy to select candidate genes based on truncating mutations is similar to that in other NGS studies (Gilissen et al. 2012). Truncating mutations introduce premature stop codons causing the dysfunction of the protein or nonsense-mediated decay whereas many missense mutations are variants of unknown significance (VUS), which means it is unclear whether they affect gene function or not. As a segregation analysis was possible in only five proband families, we were unable to prove whether the variants segregate with the phenotype and/or whether they are *de novo* events.

Based on functions and pathways related to cancers, literature review, frequency of somatic mutations in colon tumors, haploinsufficiency score, and intolerance score, *CNTN6* is as one of the 11 genes which most frequent found truncating mutations and fits into all criteria, same as *KIF26B*, *MCM3AP*, *FOCAD*, and *HSPH1*.

*CNTN6*, or contactin6, is located on chromosome 3p26.3. It is a well-known gene involved in central nervous system development and functions as an axonal guidance during development maintenance of synaptic connections in adults. It is also involved in cell adhesion and the Notch signaling pathway, which is an essential pathway for maintaining the stem cell population as well as regulating cell lineage differentiation in the normal intestinal mucosa. Recently, Smith et al. (2013) identified a truncating mutation in *NOTCH3* in one of 50 CRC patients. So far, there have been no publications of *CNTN6* as being related to CRC but a few studies have reported an association of *CNTN6* with other cancers. Manderson et

al. (2009) detected LOH within chromosome 3p25.3-pter, including *CNTN6*, in epithelial ovarian cancers. Rokman et al. (2005) reported that chromosome 3p25-p26 is a susceptibility locus for prostate cancer. However, these authors cannot observe an exonic mutation or a change of the expression level of *CNTN6*. Further studies are needed to determine whether mutations of genes involved in the Notch signaling pathway have a significant causative role in adenomatous polyposis.

*KIF26B*, or Kinesin family member 26B, is a member of the kinesin superfamily proteins (KIFs). It is known to be involved in the regulation of cell-cell adhesion. Several studies reported that abnormal expressions of different kinesins play a key role in development or progression of human cancers including CRC, partly by disturbing mitosis. Wang et al. (2013) reported that *KIF26B* is overexpressed in breast cancer tissue. Krepschi et al. (2012a) found a deletion in an early-onset familial case of breast cancer. In our study, *KIF26B* was affected by a partial duplication in one patient. However, the predominant mutation type in our cohort and in colorectal tumors is missense mutation. Germline missense mutations are outside the functional domain and cluster in exons 3 and 12. Nevertheless, this might be due to the length of the exons rather than pointing to mutational hotspots. Taken together, these results are not completely consistent and the impact of the different mutations has yet to be explored further.

*MCM3AP*, acetylating minichromosome maintenance 3 (MCM3), or *GANP*, is an essential human DNA replication protein. It is a potential natural inhibitor of the initiation of DNA replication and, thus, functions to ensure the stability of human genomic DNA (Takei et al. 2002). MCM3 acetylation has been suggested to be a novel pathway regulating DNA replication. We identified a partial duplication of the 3' part of the gene, a frameshift mutation at the beginning, and two missense variants in the 3' and 5' parts, which are predicted to be deleterious and are compatible with the expectation of loss-of-function mutations.

Another interesting candidate is *FOCAD*, previously known as *KIAA1797*. *FOCAD* is a large and highly conserved gene and has been reported in a few cancer types. So far, the germline deletions described in patients with unexplained familial tumor syndromes are not identical but partly overlap. Venkatachalam et al. (2011) found a heterozygous germline deletion (158 kb), which encompassed *FOCAD* exons 4-21 in one of 41 early-onset CRC patients. Krepschi et al. (2012a) identified a germline deletion CNV (136 kb), which encompassed *FOCAD* exons 3-13 in a patient with *BRCA1/BRCA2* mutation-negative early-onset (38 yrs) breast cancer. Recently, *FOCAD* was characterized as a novel component of the focal adhesion complex with tumor suppressor function, which was found to be disrupted by translocations or heterozygous or homozygous deletions in around half of glioblastomas (Brockschmidt et al. 2012). Focal adhesions influence growth control and are involved in cellular processes such as motility, proliferation, and differentiation (Dubash et al. 2009). Our study found a deletion of exons 20-30 and two truncating point mutations in the middle part

of the gene. These findings strengthen the pathogenic role in cancer predisposition of *FOCAD* as it might be a predisposing factor for a variety of different tumors.

*HSPH1* seems to be a haploinsufficient gene intolerant to variations. This is supported by the very low number (only 1) of truncating mutations reported in the EVS database. *HSPH1* belongs to the heat shock proteins (HSPs), which ensure the correct conformation of cellular proteins, and during stress, promote cell survival by maintaining protein homeostasis and inhibiting apoptosis (Lang et al. 2012). It suppresses H<sub>2</sub>O<sub>2</sub>-induced apoptosis by suppression of p38 MAPK signaling. Dai et al. (2012) reported that overexpression of HSPs in cancer cells reinforces oncogenic events. However, the pathogenicity of the two *HSPH1* variants identified in our patient remains unclear. The patient presented an attenuated colorectal phenotype and a family history consistent with autosomal recessive inheritance (parents healthy, his brother had few colon polyps, his sister was diagnosed with breast cancer at 45 years of age). Unfortunately, neither DNA nor RNA are available to confirm compound heterozygosity and aberrant splicing.

For a whole gene duplication, it is assumed that a gain-of-function mutation of the gene would lead to tumorigenesis. Generally, a truncating mutation of the gene would lead to loss-of-function whereas specific missense mutations are more likely to increase the function of the gene. For example, a missense mutation of *KRAS* oncogene, which change an amino acid, p.G12D, can promote the activation of RAS protein and leads pancreatic tumor (Pylayeva-Gupta et al. 2011). Regarding this assumption, we looked for missense mutations of whole gene duplication candidates.

In addition to the missense mutation in *CTNNB1*, we identified a missense mutation in *PTPN18*, which was found to be affected by a whole gene duplication in the CNV analysis. *PTPN18* belongs to the protein tyrosine phosphatase (PTP) family. PTPs regulate a variety of cellular processes including cell growth, differentiation, the mitotic cycle, and oncogenic transformation. *PTPN18* regulates HER2, a member of the epidermal growth factor receptor (EGFR) family of receptor tyrosine kinases. Overexpression of *PTPN18* inhibits HER2 activity in breast cancer (Gensler et al. 2004; Lucci et al. 2010). The frequency of somatic mutations reported in COSMIC is very low whereas the haploinsufficiency score and intolerance score show that *PTPN18* is likely to be haplosufficient and tolerated to variants.

Gain of function of *PTNP18* seems to inhibit tumor progression which is in contrast to concept of oncogenes. *PTPN18* has been reported an association with not only breast cancer but also thyroid cancer (Guimaraes et al. 2006). However, there has been no report of *PTPN18* being related to CRC so far. *PTPN18* is rather a TSG than an oncogene, correspondingly, whole gene duplications and missense mutations do not represent the typical mutation spectrum of the TSG. Based on these pieces of information, it is unlikely to be a causative gene initiating adenomatous polyposis.

Loss of function of TSG is a recessive pattern and known as “two-hit” hypothesis (Knudson 2001). Therefore, to prove the tumor suppressor activity of the candidate TSGs, identification of somatic mutation (second hit) could be helpful. This study performed somatic mutation analysis in five putative TSGs and found a missense somatic mutation in *EPHB4*, which predicted to be deleterious by in-silico tools, in the patient who carry a duplication CNV which partially involved *EPHB4* (exon 13-17 and 3'UTR). *EPHB4* is a candidate TSG (Dopeso et al. 2009; Ronsch et al. 2011), involved in angiogenesis pathways and related to colorectal cancer (Guijarro-Munoz et al. 2013). Our finding is in line with the previous studies that *EPHB4* has a tumor suppressor activity.

This thesis provides support for the functional relevance of novel predisposing genes in unexplained adenomatous polyposis. To confirm the causality and phenotype spectrum of the predisposing genes, a larger sample size to identify recurrent mutations is needed as looking for mutations is less complicated than performing functional tests of each candidate gene.

### 5.3. Limitations of the study

Similar to other studies using a genome-wide CNV profiling approach, our work has some limitations. Since we used only one calling algorithm, we were unable to determine the sensitivity of false negative callings. Since not all protein-coding exons are equally covered by SNPs, a few CNVs might not be identified due to low SNP coverage of the region.

Due to the stringent filtering process, some relevant genes might have been missed, in particular more frequent CNVs acting as moderately penetrant risk factors. Additionally, intronic CNVs, which may cause aberrant splicing and CNVs affecting regulatory regions, were excluded; however, in the context of Mendelian disease, the etiological impact of mutations in these regions is expected to be smaller compared to mutations in the coding regions.

The parents of patients with attenuated polyposis are often deceased or inaccessible, and/or the colorectal phenotype cannot be clarified reliably, so we decided against a trio setting. Thus, the *de novo* occurrence of a CNV could not be proven systematically because DNA of patients' parents was usually unavailable. However, although a high selection against *de novo* CNVs has been assumed in other phenotypes (Itsara et al. 2010; Stadler et al. 2012), the rather attenuated clinical presentation in our patient cohort argues against strong purifying selection with a high frequency of *de novo* events.

## 6. SUMMARY

In up to 50% of families with clinically verified adenomatous polyposis no germline mutations in the established genes *APC* and *MUTYH* can be identified during routine diagnostics although the presence of high numbers of colorectal adenomas strongly argues for an underlying genetic cause, either as a monogenic or genetically complex trait.

Therefore, the aim of this study was (i) to identify cryptic germline mutations in the *APC* gene which were not detected by routine diagnostics; (ii) to identify novel causative genes of adenomatous polyposis by a genome-wide SNP-array based CNV analysis, and (iii) to further evaluate the pathogenic relevance of the candidate genes by additional experiments (segregation analysis, expression analysis, screening for germline point mutations, examination of tumor tissue to identify somatic mutations).

Firstly, a functional study at the mRNA (transcript) level was carried out to look for deep intronic *APC* mutations. We identified aberrant transcript patterns in 8 (6%) of 125 unrelated patients. Five of them carried a founder germline mutation in intron 4 and three patients showed germline point mutations in intron 10, which lead to the inclusion of a pseudoexon 4 and a pseudoexon 10 on transcript level. The pseudoexons are predicted to result in frameshift mutations and premature stop codons. The mutations segregated with the disease in those families where affected relatives could be examined. Based on the results of these experiments, the germline mutations were regarded as disease-causing. Thus, a few deep intronic mutations contribute substantially to the *APC* mutation spectrum and cDNA analysis and/or target sequencing of intronic regions should be considered as an additional mutation discovery approach in polyposis patients.

To uncover novel causative genes in patients with unexplained adenomatous polyposis, a genome-wide analysis of germline copy number variants (CNV) using high-resolution SNP arrays was performed in 221 unrelated, well characterized *APC* and *MUTYH* mutation negative German patients. Putative CNVs were filtered according to stringent criteria, compared with those of 531 population-based German controls, and validated by qPCR. 125 unique rare germline CNVs in 93 (42%) of 221 patients were identified. These CNVs involved 68 deleted and 168 duplicated genes. The vast majority of patients harbor one CNV only.

A segregation analysis was conducted for nine CNVs, however, no segregation with the phenotype was observed, indicating either a lack of causal relevance for the phenotype or, in particular for the two CNVs found two times each, a low to moderate penetrance. Two CNVs in or nearby established polyposis-related genes (*CTNNB1*, *MUTYH*) seemed to be interesting, however, expression analysis on RNA level showed no significant difference in the expression levels of the two patients compared to controls.



To further evaluate the pathogenic relevance of the candidate genes, additional filtering and prioritization steps on gene level including expression analysis in cDNA from human colon tissue, network analysis, enrichment analyses of genes and pathways, and data mining were performed. Ninety-eight candidate genes remained, 32 of which showed molecular and cellular functions related to tumorigenesis. Assessing the functional relevance of mutations in these 32 genes by using the frequency of somatic mutations in colorectal tumors, and two functional scores (intolerance score, haploinsufficiency score), resulted in the selection of *ARHGAP5*, *CNTN6*, *EPHB4*, *IQGAP1*, *KIF26B*, *MCM3AP*, *NFATC1*, and *XRN1* as the most convincing candidates.

To further explore the clinical relevance of the candidate genes in the absence of recurrent alterations and lack of segregation information, a germline point mutation analysis of all candidates was performed in a validation cohort using a targeted next generation sequencing (NGS) approach. Fifteen rare heterozygous truncating point mutations in 11 genes (*CNTN6*, *FOCAD*, *HEXDC*, *HSPH1*, *KIF26B*, *MCM3AP*, *PXDNL*, *TESK2*, *ULK4*, *YBEY*, and *ZNF471*) were identified in 15 patients. In these 11 genes, we found additional 27 rare missense mutations which were predicted to be deleterious. *CNTN6* and *FOCAD* showed different truncating mutations in more than one patient whereas *KIF26B* has the highest frequency of potential deleterious mutations overall. The causative relevance of the two suggested tumor suppressor genes, *FOCAD* and *EPHB4*, was further underscored by the detection of somatic point mutations (“second hits”) in tumor tissue of the patients.

By integrating all results and recent studies of early-onset colorectal and breast cancer, we selected *CNTN6*, *EPHB4*, *KIF26B*, *MCM3AP*, *FOCAD*, and *HSPH1* as the most convincing predisposing genes for colorectal adenomatous polyposis. In addition, in the canonical Wnt pathway oncogene *CTNNB1* ( $\beta$ -catenin), two potential gain-of-function mutations were found.

This thesis identified a group of rarely affected genes which are likely to predispose to colorectal adenoma formation and confirmed previously published candidates for tumor predisposition as etiologically relevant. Based on this work, rare CNVs are likely to contribute to the hereditary risk for colorectal tumors. However, the sporadic disease manifestation in most families, the often incomplete segregation of the genetic alterations, and the occurrence in the context of different tumor types indicate that at least some of the candidate genes act rather as moderate penetrant risk factors than highly penetrant mutations. The identification of six patients with di- or trigenic alterations is consistent with the assumption of a more oligogenic etiology. This approach generated no evidence for a high frequency of recessive subtypes. Similar to recent studies, the vast majority of the rare CNVs was non-recurrent. Our analysis demonstrated that the underlying genetic factors of unexplained colorectal polyposis are likely to be very heterogeneous, which makes clinical validation challenging. To further characterize the functional relevance of the selected genes, international collaborations with large patient cohorts and functional studies are needed.

## 7. OUTLOOK/PERSPECTIVE

The transcript analysis of the APC gene pointed to the existence of deep intronic germline mutations, in particular a putative mutational hotspots in intron 10 and a founder mutation in intron 4. To further explore the frequency and distribution of pathogenic intronic mutations it will be interesting to examine additional patient cohorts and other polyposis forms. An extension of the routine germline mutation screening in FAP patients by sequencing of the regions around the identified hotspot and founder mutations should be considered a reasonable procedure.

To further validate the causative relevance of the identified candidate genes of the CNV analysis it is important to identify recurrent mutations of the same mutation type in the respective genes. To achieve this, large collaborative studies including patient cohorts from several groups and the submission of all identified variants in locus-specific databases are needed. The already existing collaborations with several relevant groups are the basis to initiate such projects.

In this study, we focus only on rare CNVs covering protein coding regions. In future studies, rare CNVs in intergenic and intronic regions might be considered as well. Such locations might be related to regulatory regions or might result in aberrant splicing, however, the consequences of such CNVs are difficult to predict.

We used an approach assuming a monogenic bases of the polyposis disease. However, a major group of the phenotype might be better explained by oligogenic or genetically complex etiology. To address this, the data can be used to perform a CNV-based GWAS and a CNV-based burden analysis. In addition, a SNP-based GWAS to look for low penetrance variants and a replication study have been performed recently, the statistical analyses are still ongoing.

To identify genomic regions with potential causative disease loci it is also possible to look for shared haplotypes among the patients. The haplotype sharing statistic (HSS) compares the length of shared haplotypes between cases and controls. This approach might be even more powerful when large patient cohorts with the same ethnical background are included.

The CNV approach identifies only large genomic deletions and duplications. To identify the whole mutational spectrum, a high-throughput sequencing of the coding regions of the genome (exome sequencing) or even the whole genome is needed. Currently, exome sequencing is regarded to be the most powerful tool to identify the causative genes of the still unexplained monogenic conditions. We have already started collaborative projects of exome sequencing in around 100 patients of the study cohort.

The most interesting genes identified by these projects must be subject of functional studies to characterize the pathophysiological consequences of the mutations. Such studies may include expression analyses, interaction and pathway analyses performed in cell culture experiments, and mouse models.

## 8. REFERENCES

- Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, Hodges AK, Davies DR, David SS, Sampson JR, Cheadle JP (2002) Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* 30: 227-32.
- Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* 12: 363-76.
- Aretz S (2010) The differential diagnosis and surveillance of hereditary gastrointestinal polyposis syndromes. *Dtsch Arztebl Int* 107: 163-73.
- Aretz S, Genuardi M, Hes FJ (2013) Clinical utility gene card for: MUTYH-associated polyposis (MAP), autosomal recessive colorectal adenomatous polyposis, multiple colorectal adenomas, multiple adenomatous polyps (MAP) - update 2012. *Eur J Hum Genet* 21.
- Aretz S, Stienen D, Friedrichs N, Stemmler S, Uhlhaas S, Rahner N, Propping P, Friedl W (2007a) Somatic APC mosaicism: a frequent cause of familial adenomatous polyposis (FAP). *Hum Mutat* 28: 985-92.
- Aretz S, Stienen D, Uhlhaas S, Loff S, Back W, Pagenstecher C, McLeod DR, Graham GE, Mangold E, Santer R, Propping P, Friedl W (2005) High proportion of large genomic STK11 deletions in Peutz-Jeghers syndrome. *Hum Mutat* 26: 513-9.
- Aretz S, Stienen D, Uhlhaas S, Stolte M, Entius MM, Loff S, Back W, Kaufmann A, Keller KM, Blaas SH, Siebert R, Vogt S, Spranger S, Holinski-Feder E, Sunde L, Propping P, Friedl W (2007b) High proportion of large genomic deletions and a genotype phenotype update in 80 unrelated families with juvenile polyposis syndrome. *J Med Genet* 44: 702-9.
- Aretz S, Uhlhaas S, Goergens H, Siberg K, Vogel M, Pagenstecher C, Mangold E, Caspari R, Propping P, Friedl W (2006) MUTYH-associated polyposis: 70 of 71 patients with biallelic mutations present with an attenuated or atypical phenotype. *Int J Cancer* 119: 807-14.
- Aretz S, Uhlhaas S, Sun Y, Pagenstecher C, Mangold E, Caspari R, Moslein G, Schulmann K, Propping P, Friedl W (2004) Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Hum Mutat* 24: 370-80.
- Azuma M, Shi M, Danenberg KD, Gardner H, Barrett C, Jacques CJ, Sherod A, Iqbal S, El-Khoueiry A, Yang D, Zhang W, Danenberg PV, Lenz HJ (2007) Serum lactate dehydrogenase levels and glycolysis significantly correlate with tumor VEGFA and VEGFR expression in metastatic CRC patients. *Pharmacogenomics* 8: 1705-13.
- Bagnall RD, Waseem NH, Green PM, Colvin B, Lee C, Giannelli F (1999) Creation of a novel donor splice site in intron 1 of the factor VIII gene leads to activation of a 191 bp cryptic exon in two haemophilia A patients. *Br J Haematol* 107: 766-71.
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. *Science* 297: 1003-7.
- Barraclough J, Hodgkinson C, Hogg A, Dive C, Welman A (2007) Increases in c-Yes expression level and activity promote motility but not proliferation of human colorectal carcinoma cells. *Neoplasia* 9: 745-54.
- Beroud C, Carrie A, Beldjord C, Deburgrave N, Llense S, Carelle N, Peccate C, Cuisset JM, Pandit F, Carre-Pigeon F, Mayer M, Bellance R, Recan D, Chelly J, Kaplan JC,

- Leturcq F (2004) Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene. *Neuromuscul Disord* 14: 10-8.
- Bertotti A, Migliardi G, Galimi F, Sassi F, Torti D, Isella C, Cora D, Di Nicolantonio F, Buscarino M, Petti C, Ribero D, Russolillo N, Muratore A, Massucco P, Pisacane A, Molinaro L, Valtorta E, Sartore-Bianchi A, Risio M, Capussotti L, Gambacorta M, Siena S, Medico E, Sapino A, Marsoni S, Comoglio PM, Bardelli A, Trusolino L (2011) A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discov* 1: 508-23.
- Brault LS, Cooper CA, Famula TR, Murray JD, Penedo MC (2011) Mapping of equine cerebellar abiotrophy to ECA2 and identification of a potential causative mutation affecting expression of MUTYH. *Genomics* 97: 121-9.
- Braun R, Buetow K (2011) Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. *PLoS Genet* 7: e1002101.
- Brewer C, Holloway S, Zawalnyski P, Schinzel A, FitzPatrick D (1999) A chromosomal duplication map of malformations: regions of suspected haplo- and triplolethality--and tolerance of segmental aneuploidy--in humans. *Am J Hum Genet* 64: 1702-8.
- Brim H, Lee E, Abu-Asab MS, Chaouchi M, Razjouyan H, Namin H, Goel A, Schaffer AA, Ashktorab H (2012) Genomic aberrations in an African American colorectal cancer cohort reveals a MSI-specific profile and chromosome X amplification in male patients. *PLoS One* 7: e40392.
- Brockschmidt A, Trost D, Peterziel H, Zimmermann K, Ehrler M, Grassmann H, Pfenning PN, Waha A, Wohlleber D, Brockschmidt FF, Jugold M, Hoischen A, Kalla C, Seifert G, Knolle PA, Latz E, Hans VH, Wick W, Pfeifer A, Angel P, Weber RG (2012) KIAA1797/FOCAD encodes a novel focal adhesion protein with tumour suppressor function in gliomas. *Brain* 135: 1027-41.
- Buizer-Voskamp JE, Muntjewerff JW, Strengman E, Sabatti C, Stefansson H, Vorstman JA, Ophoff RA (2011) Genome-wide analysis shows increased frequency of copy number variation deletions in Dutch schizophrenia patients. *Biol Psychiatry* 70: 655-62.
- Burset M, Seledtsov IA, Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* 28: 4364-75.
- Cardullo RA, Agrawal S, Flores C, Zamecnik PC, Wolf DE (1988) Detection of nucleic acid hybridization by nonradiative fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A* 85: 8790-4.
- Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res* 31: 3568-71.
- Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 39: S16-21.
- Chayka O, Corvetta D, Dews M, Caccamo AE, Piotrowska I, Santilli G, Gibson S, Sebire NJ, Himoudi N, Hogarty MD, Anderson J, Bettuzzi S, Thomas-Tikhonenko A, Sala A (2009) Clusterin, a haploinsufficient tumor suppressor gene in neuroblastomas. *J Natl Cancer Inst* 101: 663-77.
- Chen W, Yuan L, Cai Y, Chen X, Chi Y, Wei P, Zhou X, Shi D (2013) Identification of chromosomal copy number variations and novel candidate loci in hereditary nonpolyposis colorectal cancer with mismatch repair proficiency. *Genomics* 102: 27-34.
- Chen X, Halberg RB, Ehrhardt WM, Torrealba J, Dove WF (2003) Clusterin as a biomarker in murine and human intestinal neoplasia. *Proc Natl Acad Sci U S A* 100: 9530-5.
- Chillon M, Dork T, Casals T, Gimenez J, Fonknechten N, Will K, Ramos D, Nunes V, Estivill X (1995) A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+1.6kbA-->G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. *Am J Hum Genet* 56: 623-9.

- Christie M, Jorissen RN, Mouradov D, Sakthianandeswaren A, Li S, Day F, Tsui C, Lipton L, Desai J, Jones IT, McLaughlin S, Ward RL, Hawkins NJ, Ruszkiewicz AR, Moore J, Burgess AW, Busam D, Zhao Q, Strausberg RL, Simpson AJ, Tomlinson IP, Gibbs P, Sieber OM (2013) Different APC genotypes in proximal and distal sporadic colorectal cancers suggest distinct WNT/beta-catenin signalling thresholds for tumourigenesis. *Oncogene* 32: 4675-82.
- Clendenning M, Buchanan DD, Walsh MD, Nagler B, Rosty C, Thompson B, Spurdle AB, Hopper JL, Jenkins MA, Young JP (2011) Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer* 10: 297-301.
- Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 35: 2013-25.
- Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 38: 75-81.
- Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464: 704-12.
- Cook EH, Jr., Scherer SW (2008) Copy-number variations associated with neuropsychiatric conditions. *Nature* 455: 919-23.
- Cooper GM, Nickerson DA, Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* 39: S22-9.
- Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* 40: 1199-203.
- Dabrowska M, Skoneczny M, Rode W (2011) Functional gene expression profile underlying methotrexate-induced senescence in human colon cancer cells. *Tumour Biol* 32: 965-76.
- Dai C, Dai S, Cao J (2012) Proteotoxic stress of cancer: implication of the heat-shock response in oncogenesis. *J Cell Physiol* 227: 2982-7.
- Daley D (2010) The identification of colon cancer susceptibility genes by using genome-wide scans. *Methods Mol Biol* 653: 3-21.
- Dauber A, Yu Y, Turchin MC, Chiang CW, Meng YA, Demerath EW, Patel SR, Rich SS, Rotter JI, Schreiner PJ, Wilson JG, Shen Y, Wu BL, Hirschhorn JN (2011) Genome-wide association of copy-number variation reveals an association between short stature and the presence of low-frequency genomic deletions. *Am J Hum Genet* 89: 751-9.
- De Klein A, Riegman PH, Bijlsma EK, Helderdoorn A, Muijtjens M, den Bakker MA, Avezaat CJ, Zwarthoff EC (1998) A G-->A transition creates a branch point sequence and activation of a cryptic exon, resulting in the hereditary disorder neurofibromatosis 2. *Hum Mol Genet* 7: 393-8.
- de Smith AJ, Walters RG, Froguel P, Blakemore AI (2008) Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease. *Cytogenet Genome Res* 123: 17-26.
- de Voer RM, Geurts van Kessel A, Weren RD, Ligtenberg MJ, Smeets D, Fu L, Vreede L, Kamping EJ, Verwiel ET, Hahn MM, Ariaans M, Spruijt L, van Essen T, Houge G, Schackert HK, Sheng JQ, Venselaar H, van Ravenswaaij-Arts CM, van Krieken JH, Hoogerbrugge N, Kuiper RP (2013) Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are risk factors for colorectal cancer. *Gastroenterology* 145: 544-7.

- Dehainault C, Michaux D, Pages-Berhouet S, Caux-Moncoutier V, Doz F, Desjardins L, Couturier J, Parent P, Stoppa-Lyonnet D, Gauthier-Villars M, Houdayer C (2007) A deep intronic mutation in the RB1 gene leads to intronic sequence exonisation. *Eur J Hum Genet* 15: 473-7.
- Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res* 38: e105.
- Dihlmann S, Gebert J, Siermann A, Herfarth C, von Knebel Doeberitz M (1999) Dominant negative effect of the APC1309 mutation: a possible explanation for genotype-phenotype correlations in familial adenomatous polyposis. *Cancer Res* 59: 1857-60.
- Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AI, London WB, Shaikh TH, Bradfield J, Grant SF, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459: 987-91.
- Dopeso H, Mateo-Lozano S, Mazzolini R, Rodrigues P, Lagares-Tena L, Ceron J, Romero J, Esteves M, Landolfi S, Hernandez-Losa J, Castano J, Wilson AJ, Ramon y Cajal S, Mariadason JM, Schwartz S, Jr., Arango D (2009) The receptor tyrosine kinase EPHB4 has tumor suppressor activities in intestinal tumorigenesis. *Cancer Res* 69: 7430-8.
- Dubash AD, Menold MM, Samson T, Boulter E, Garcia-Mata R, Doughman R, Burridge K (2009) Chapter 1. Focal adhesions: new angles on an old structure. *Int Rev Cell Mol Biol* 277: 1-65.
- Eldai H, Periyasamy S, Al Qarni S, Al Rodayyan M, Muhammed Mustafa S, Deeb A, Al Sheikh E, Afzal Khan M, Johani M, Yousef Z, Aziz MA (2013) Novel genes associated with colorectal cancer are revealed by high resolution cytogenetic analysis in a patient specific manner. *PLoS One* 8: e76251.
- Emmert-Streib F, Glazko GV (2011) Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS Comput Biol* 7: e1002053.
- Engels H, Wohlleber E, Zink A, Hoyer J, Ludwig KU, Brockschmidt FF, Wieczorek D, Moog U, Hellmann-Mersch B, Weber RG, Willatt L, Kreiss-Nachtsheim M, Firth HV, Rauch A (2009) A novel microdeletion syndrome involving 5q14.3-q15: clinical and molecular cytogenetic characterization of three patients. *Eur J Hum Genet* 17: 1592-9.
- Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics* 23: 257-8.
- Fanciulli M, Petretto E, Aitman TJ (2010) Gene copy number variation and common human disease. *Clin Genet* 77: 201-13.
- Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. *Genes Dev* 17: 419-37.
- Fearon ER (2011) Molecular genetics of colorectal cancer. *Annu Rev Pathol* 6: 479-507.
- Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet* 15 Spec No 1: R57-66.
- Fischer H, Salahshor S, Stenling R, Bjork J, Lindmark G, Iselius L, Rubio C, Lindblom A (2001) COL11A1 in FAP polyps and in sporadic colorectal tumors. *BMC Cancer* 1: 17.
- Flora M, Piana S, Bassano C, Bisagni A, De Marco L, Ciarrocchi A, Tagliavini E, Gardini G, Tamagnini I, Banzi C, Bisagni G (2012) Epidermal growth factor receptor (EGFR) gene copy number in colorectal adenoma-carcinoma progression. *Cancer Genet* 205: 630-5.
- Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. *Genome Res* 16: 949-61.

- Friedl W, Aretz S (2005) Familial adenomatous polyposis: experience from a study of 1164 unrelated german polyposis patients. *Hered Cancer Clin Pract* 3: 95-114.
- Friedl W, Lamberti C (2001) Familiäre adenomatöse Polyposis. *Hereditäre Tumorerkrankungen*. Springer-Verlag Berlin, Heidelberg, pp 303-325
- Galiatsatos P, Foulkes WD (2006) Familial adenomatous polyposis. *Am J Gastroenterol* 101: 385-98.
- Gardner EJ, Richards RC (1953) Multiple cutaneous and subcutaneous lesions occurring simultaneously with hereditary polyposis and osteomatosis. *Am J Hum Genet* 5: 139-47.
- Gensler M, Buschbeck M, Ullrich A (2004) Negative regulation of HER2 signaling by the PEST-type protein-tyrosine phosphatase BDP1. *J Biol Chem* 279: 12110-6.
- Gianni D, Bohl B, Courtneidge SA, Bokoch GM (2008) The involvement of the tyrosine kinase c-Src in the regulation of reactive oxygen species generation mediated by NADPH oxidase-1. *Mol Biol Cell* 19: 2984-94.
- Gianni D, Taulet N, DerMardirossian C, Bokoch GM (2010) c-Src-mediated phosphorylation of NoxA1 and Tks4 induces the reactive oxygen species (ROS)-dependent formation of functional invadopodia in human colon cancer cells. *Mol Biol Cell* 21: 4287-98.
- Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 20: 490-7.
- Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, Sunyaev S (2013) Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* 14: 460-70.
- Goldstein DR, Dudoit S, Speed TP (2001) Power and robustness of a score test for linkage analysis of quantitative traits using identity by descent data on sib pairs. *Genet Epidemiol* 20: 415-31.
- Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, Ley TJ (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* 3: e3.
- Greenman CD (2012) Cancer. Haploinsufficient gene selection in cancer. *Science* 337: 47-8.
- Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, Joslyn G, Stevens J, Spirio L, Robertson M, et al. (1991) Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* 66: 589-600.
- Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements. *Pathogenetics* 1: 4.
- Guijarro-Munoz I, Sanchez A, Martinez-Martinez E, Garcia JM, Salas C, Provencio M, Alvarez-Vallina L, Sanz L (2013) Gene expression profiling identifies EPHB4 as a potential predictive biomarker in colorectal cancer patients treated with bevacizumab. *Med Oncol* 30: 572.
- Guimaraes GS, Latini FR, Camacho CP, Maciel RM, Dias-Neto E, Cerutti JM (2006) Identification of candidates for tumor-specific alternative splicing in the thyroid. *Genes Chromosomes Cancer* 45: 540-53.
- Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, Doucet D, Milewski M, Yang R, Siegmund C, Haas J, Zhou L, Oliphant A, Fan JB, Barnard S, Chee MS (2004) Decoding randomly ordered DNA arrays. *Genome Res* 14: 870-7.
- Gylfe AE, Katainen R, Kondelin J, Tanskanen T, Cajuso T, Hanninen U, Taipale J, Taipale M, Renkonen-Sinisalo L, Jarvinen H, Mecklin JP, Kilpivaara O, Pitkanen E, Vahteristo P, Tuupanen S, Karhu A, Aaltonen LA (2013) Eleven candidate susceptibility genes for common familial colorectal cancer. *PLoS Genet* 9: e1003876.
- Hamamoto R, Furukawa Y, Morita M, Iimura Y, Silva FP, Li M, Yagyu R, Nakamura Y (2004) SMYD3 encodes a histone methyltransferase involved in the proliferation of cancer cells. *Nat Cell Biol* 6: 731-40.



- Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. *Nat Rev Genet* 10: 551-64.
- Hayashi H, Nabeshima K, Aoki M, Hamasaki M, Enatsu S, Yamauchi Y, Yamashita Y, Iwasaki H (2010) Overexpression of IQGAP1 in advanced colorectal cancer correlates with poor prognosis-critical role in tumor invasion. *Int J Cancer* 126: 2563-74.
- Hayashi Y, Widjono YW, Ohta K, Hanioka K, Obayashi C, Itoh K, Imai Y, Itoh H (1994) Expression of EGF, EGF-receptor, p53, v-erb B and ras p21 in colorectal neoplasms by immunostaining paraffin-embedded tissues. *Pathol Int* 44: 124-30.
- Heinen CD (2010) Genotype to phenotype: analyzing the effects of inherited mutations in colorectal cancer families. *Mutat Res* 693: 32-45.
- Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A (2009) Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* 41: 424-9.
- Hermes S, Hofmann A, Priebe L, Degenhardt F, Mühleisen T, Nöthen M, Cichon S, Hoffmann P (2013) Comparison of copy number variation (CNV) calling performance in large numbers of technical replicate SNP array data using three different, widely-used CNV calling algorithms European Human Genetics Conference 2013
- Hes FJ, Nielsen M, Bik EC, Konvalinka D, Wijnen JT, Bakker E, Vasen HF, Breuning MH, Tops CM (2008) Somatic APC mosaicism: an underestimated cause of polyposis coli. *Gut* 57: 71-6.
- Hes FJ, Ruano D, Nieuwenhuis M, Tops CM, Schrumpf M, Nielsen M, Huijts PE, Wijnen JT, Wagner A, Gomez Garcia EB, Sijmons RH, Menko FH, Letteboer TG, Hoogerbrugge N, Harryvan J, Kampman E, Morreau H, Vasen HF, van Wezel T (2014) Colorectal cancer risk variants on 11q23 and 15q13 are associated with unexplained adenomatous polyposis. *J Med Genet* 51: 55-60.
- Highsmith WE, Burch LH, Zhou Z, Olsen JC, Boat TE, Spock A, Gorvoy JD, Quittel L, Friedman KJ, Silverman LM, et al. (1994) A novel mutation in the cystic fibrosis gene in patients with pulmonary disease but normal sweat chloride concentrations. *N Engl J Med* 331: 974-80.
- Homolova K, Zavadakova P, Doktor TK, Schroeder LD, Kozich V, Andresen BS (2010) The deep intronic c.903+469T>C mutation in the MTRR gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cblE type of homocystinuria. *Hum Mutat* 31: 437-44.
- Hoyer J, Ekici AB, Ende S, Popp B, Zweier C, Wiesener A, Wohlleber E, Dufke A, Rossier E, Petsch C, Zweier M, Gohring I, Zink AM, Rappold G, Schrock E, Wieczorek D, Riess O, Engels H, Rauch A, Reis A (2012) Haploinsufficiency of ARID1B, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability. *Am J Hum Genet* 90: 565-72.
- Huang CW, Tsai HL, Chen YT, Huang CM, Ma CJ, Lu CY, Kuo CH, Wu DC, Chai CY, Wang JY (2013) The prognostic values of EGFR expression and KRAS mutation in patients with synchronous or metachronous metastatic colorectal cancer. *BMC Cancer* 13: 599.
- Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* 6: e1001154.
- International-HapMap-Consortium (2005) A haplotype map of the human genome. *Nature* 437: 1299-320.
- Ioannidis JP, Castaldi P, Evangelou E (2010) A compendium of genome-wide associations for cancer: critical synopsis and reappraisal. *J Natl Cancer Inst* 102: 846-58.
- Ionita-Laza I, Ottman R (2011) Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 189: 1061-8.

- Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C (2009) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics* 93: 22-6.
- Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84: 148-61.
- Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE (2010) De novo rates and selection of large copy number variation. *Genome Res* 20: 1469-81.
- Jager M, Schoberth A, Ruf P, Hess J, Hennig M, Schmalfeldt B, Wimberger P, Strohlein M, Theissen B, Heiss MM, Lindhofer H (2012) Immunomonitoring results of a phase II/III study of malignant ascites patients treated with the trifunctional antibody catumaxomab (anti-EpCAM x anti-CD3). *Cancer Res* 72: 24-32.
- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003.
- Jasperson KW, Tuohy TM, Neklason DW, Burt RW (2010) Hereditary and familial colon cancer. *Gastroenterology* 138: 2044-58.
- Jiang L, Jiang J, Yang J, Liu X, Wang J, Wang H, Ding X, Liu J, Zhang Q (2013) Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics* 14: 131.
- Jones S, Emmerson P, Maynard J, Best JM, Jordan S, Williams GT, Sampson JR, Cheadle JP (2002) Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations. *Hum Mol Genet* 11: 2961-7.
- Karameris A, Kanavaros P, Aninos D, Gorgoulis V, Mikou G, Rokas T, Niotis M, Kalogeropoulos N (1993) Expression of epidermal growth factor (EGF) and epidermal growth factor receptor (EGFR) in gastric and colorectal carcinomas. An immunohistological study of 63 cases. *Pathol Res Pract* 189: 133-7.
- Kaufmann A, Vogt S, Uhlhaas S, Stienen D, Kurth I, Hameister H, Mangold E, Kotting J, Kaminsky E, Propping P, Friedl W, Aretz S (2009) Analysis of rare APC variants at the mRNA level: six pathogenic mutations and literature review. *J Mol Diagn* 11: 131-9.
- Kim H, Watkinson J, Varadan V, Anastassiou D (2010) Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med Genomics* 3: 51.
- King K, Flinter FA, Nihalani V, Green PM (2002) Unusual deep intronic mutations in the COL4A5 gene cause X linked Alport syndrome. *Hum Genet* 111: 548-54.
- Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hedge P, McKechnie D, et al. (1991) Identification of FAP locus genes from chromosome 5q21. *Science* 253: 661-5.
- Kleinjan DJ, van Heyningen V (1998) Position effect in human genetic disease. *Hum Mol Genet* 7: 1611-8.
- Knudson AG (2001) Two genetic hits (more or less) to cancer. *Nat Rev Cancer* 1: 157-62.
- Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318: 420-6.
- Krepischi AC, Achatz MI, Santos EM, Costa SS, Lisboa BC, Brentani H, Santos TM, Goncalves A, Nobrega AF, Pearson PL, Vianna-Morgante AM, Carraro DM, Brentani

- RR, Rosenberg C (2012a) Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* 14: R24.
- Krepischi AC, Pearson PL, Rosenberg C (2012b) Germline copy number variations and cancer predisposition. *Future Oncol* 8: 441-50.
- Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS (2010) The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* 55: 403-15.
- Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A (2010) Germline copy number variation and cancer risk. *Curr Opin Genet Dev* 20: 282-9.
- Kuiper RP, Vissers LE, Venkatachalam R, Bodmer D, Hoenselaar E, Goossens M, Haufe A, Kamping E, Niessen RC, Hogervorst FB, Gille JJ, Redeker B, Tops CM, van Gijn ME, van den Ouweland AM, Rahner N, Steinke V, Kahl P, Holinski-Feder E, Morak M, Kloor M, Stemmler S, Betz B, Hutter P, Bunyan DJ, Syngal S, Culver JO, Graham T, Chan TL, Nagtegaal ID, van Krieken JH, Schackert HK, Hoogerbrugge N, van Kessel AG, Ligtenberg MJ (2011) Recurrence and variability of germline EPCAM deletions in Lynch syndrome. *Hum Mutat* 32: 407-14.
- Kwon JM, Goate AM (2000) The candidate gene approach. *Alcohol Res Health* 24: 164-8.
- Lambertz I, Nittner D, Mestdagh P, Denecker G, Vandesompele J, Dyer MA, Marine JC (2010) Monoallelic but not biallelic loss of Dicer1 promotes tumorigenesis in vivo. *Cell Death Differ* 17: 633-41.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaanty KD, Miner TL, Delehaanty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
- Lang BJ, Nguyen L, Nguyen HC, Vieusseux JL, Chai RC, Christophi C, Fifis T, Kouspou MM, Price JT (2012) Heat stress induces epithelial plasticity and cell migration independent of heat shock factor 1. *Cell Stress Chaperones* 17: 765-78.
- Le SV, Yamaguchi DJ, McArdle CA, Tachiki K, Pisegna JR, Germano P (2002) PAC1 and PACAP expression, signaling, and effect on the growth of HCT8, human colonic tumor cells. *Regul Pept* 109: 115-25.
- Leclerc D, Cao Y, Deng L, Mikael LG, Wu Q, Rozen R (2013) Differential gene expression and methylation in the retinoid/PPARA pathway and of tumor suppressors may modify intestinal tumorigenesis induced by low folate in mice. *Mol Nutr Food Res* 57: 686-97.
- Lee C, Scherer SW (2010) The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 12: e8.
- Lee JA, Carvalho CM, Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131: 1235-47.
- Lee JA, Inoue K, Cheung SW, Shaw CA, Stankiewicz P, Lupski JR (2006) Role of genomic architecture in PLP1 duplication causing Pelizaeus-Merzbacher disease. *Hum Mol Genet* 15: 2250-65.

- Lee W, Belkhir A, Lockhart AC, Merchant N, Glaeser H, Harris EI, Washington MK, Brunt EM, Zaika A, Kim RB, El-Rifai W (2008) Overexpression of OATP1B3 confers apoptotic resistance in colon cancer. *Cancer Res* 68: 10315-23.
- Lefevre JH, Bonilla C, Colas C, Winney B, Johnstone E, Tonks S, Day T, Hutnik K, Boumertit A, Soubrier F, Midgley R, Kerr D, Parc Y, Bodmer WF (2012) Role of rare variants in undetermined multiple adenomatous polyposis and early-onset colorectal cancer. *J Hum Genet* 57: 709-16.
- Lelievre V, Meunier AC, Caigneaux E, Falcon J, Muller JM (1998) Differential expression and function of PACAP and VIP receptors in four human colonic adenocarcinoma cell lines. *Cell Signal* 10: 13-26.
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-60.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-9.
- Lieber MR (2008) The mechanism of human nonhomologous DNA end joining. *J Biol Chem* 283: 1-5.
- Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Wiklund F, Wiley K, Isaacs SD, Stattin P, Xu J, Duggan D, Carpten JD, Isaacs WB, Gronberg H, Zheng SL, Chang BL (2009) Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. *Cancer Res* 69: 2176-9.
- Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2<sup>-</sup>(Delta Delta C(T)) Method. *Methods* 25: 402-8.
- Lucci-Cordisco E, Zollino M, Baglioni S, Mancuso I, Lecce R, Gurrieri F, Crucitti A, Papi L, Neri G, Genuardi M (2005) A novel microdeletion syndrome with loss of the MSH2 locus and hereditary non-polyposis colorectal cancer. *Clin Genet* 67: 178-82.
- Lucci MA, Orlandi R, Triulzi T, Tagliabue E, Balsari A, Villa-Moruzzi E (2010) Expression profile of tyrosine phosphatases in HER2 breast cancer cells and tumors. *Cell Oncol* 32: 361-72.
- Lucito R, Suresh S, Walter K, Pandey A, Lakshmi B, Krasnitz A, Sebat J, Wigler M, Klein AP, Brune K, Palmisano E, Maitra A, Goggins M, Hruban RH (2007) Copy-number variants in patients with a strong family history of pancreatic cancer. *Cancer Biol Ther* 6: 1592-9.
- Makalowska I (2008) Comparative analysis of an unusual gene arrangement in the human chromosome 1. *Gene* 423: 172-9.
- Manderson EN, Birch AH, Shen Z, Mes-Masson AM, Provencher D, Tonin PN (2009) Molecular genetic analysis of a cell adhesion molecule with homology to L1CAM, contactin 6, and contactin 4 candidate chromosome 3p26pter tumor suppressor genes in ovarian cancer. *Int J Gynecol Cancer* 19: 513-25.
- Marenne G, Real FX, Rothman N, Rodriguez-Santiago B, Perez-Jurado L, Kogevinas M, Garcia-Closas M, Silverman DT, Chanock SJ, Genin E, Malats N (2012) Genome-wide CNV analysis replicates the association between GSTM1 deletion and bladder cancer: a support for using continuous measurement from SNP-array data. *BMC Genomics* 13: 326.
- Matsubara J, Honda K, Ono M, Sekine S, Tanaka Y, Kobayashi M, Jung G, Sakuma T, Nakamori S, Sata N, Nagai H, Ioka T, Okusaka T, Kosuge T, Tsuchida A, Shimahara M, Yasunami Y, Chiba T, Yamada T (2011) Identification of adipophilin as a potential plasma biomarker for colorectal cancer using label-free quantitative mass spectrometry and protein microarray. *Cancer Epidemiol Biomarkers Prev* 20: 2195-203.
- Mazzarelli P, Pucci S, Spagnoli LG (2009) CLU and colon cancer. The dual face of CLU: from normal to malignant phenotype. *Adv Cancer Res* 105: 45-61.

- McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM (2006) Common deletion polymorphisms in the human genome. *Nat Genet* 38: 86-92.
- McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40: 1166-74.
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9: 356-69.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-303.
- Meindl A, Hellebrand H, Wiek C, Erven V, Wappenschmidt B, Niederacher D, Freund M, Lichtner P, Hartmann L, Schaal H, Ramser J, Honisch E, Kubisch C, Wichmann HE, Kast K, Deissler H, Engel C, Muller-Myhsok B, Neveling K, Kiechle M, Mathew CG, Schindler D, Schmutzler RK, Hanenberg H (2010) Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. *Nat Genet* 42: 410-4.
- Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 11: 685-96.
- Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 16: 1215.
- Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HY, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stutz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korb JO (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470: 59-65.
- Mitra PS, Ghosh S, Zang S, Sonneborn D, Hertz-Picciotto I, Trnovec T, Palkovicova L, Sovcikova E, Ghimbovski S, Hoffman EP, Dutta SK (2012) Analysis of the toxicogenomic effects of exposure to persistent organic pollutants (POPs) in Slovakian girls: correlations between gene expression and disease risk. *Environ Int* 39: 188-99.
- Mongin C, Coulet F, Lefevre JH, Colas C, Svrcek M, Eyries M, Lahely Y, Flejou JF, Soubrier F, Parc Y (2012) Unexplained polyposis: a challenge for geneticists, pathologists and gastroenterologists. *Clin Genet* 81: 38-46.
- Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, Anaya V, Richardson R, Davis J, MacArthur DG, Sidow A, Duret L, Gerstein M, Makova KD, Marchini J, McVean G, Lunter G (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* 23: 749-61.
- Moran AE, Hunt DH, Javid SH, Redston M, Carothers AM, Bertagnolli MM (2004) Apc deficiency is associated with increased Egfr activity in the intestinal enterocytes and adenomas of C57BL/6J-Min/+ mice. *J Biol Chem* 279: 43261-72.
- Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 13: 523-36.

- Moriya M, Grollman AP (1993) Mutations in the mutY gene of *Escherichia coli* enhance the frequency of targeted G:C-->T:a transversions induced by a single 8-oxoguanine residue in single-stranded DNA. *Mol Gen Genet* 239: 72-6.
- Mullaney JM, Mills RE, Pittard WS, Devine SE (2010) Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* 19: R131-6.
- Myllykangas S, Natsoulis G, Bell JM, Ji HP (2011) Targeted sequencing library preparation by genomic DNA circularization. *BMC Biotechnol* 11: 122.
- Nabeshima K, Shimao Y, Inoue T, Kono M (2002) Immunohistochemical analysis of IQGAP1 expression in human colorectal carcinomas: its overexpression in carcinomas and association with invasion fronts. *Cancer Lett* 176: 101-9.
- Natrajan R, Mackay A, Lambros MB, Weigelt B, Wilkerson PM, Manie E, Grigoriadis A, A'Hern R, van der Groep P, Kozarewa I, Popova T, Mariani O, Turajlic S, Furney SJ, Marais R, Rodruigues DN, Flora AC, Wai P, Pawar V, McDade S, Carroll J, Stoppa-Lyonnet D, Green AR, Ellis IO, Swanton C, van Diest P, Delattre O, Lord CJ, Foulkes WD, Vincent-Salomon A, Ashworth A, Henri Stern M, Reis-Filho JS (2012) A whole-genome massively parallel sequencing analysis of BRCA1 mutant oestrogen receptor-negative and -positive breast cancers. *J Pathol* 227: 29-41.
- Nemetz N, Abad C, Lawson G, Nobuta H, Chhith S, Duong L, Tse G, Braun J, Waschek JA (2008) Induction of colitis and rapid development of colorectal tumors in mice deficient in the neuropeptide PACAP. *Int J Cancer* 122: 1803-9.
- Neumann J, Wehweck L, Maatz S, Engel J, Kirchner T, Jung A (2013) Alterations in the EGFR pathway coincide in colorectal cancer and impact on prognosis. *Virchows Arch* 463: 509-23.
- Newton K, Mallinson E, Bowen J, Laloo F, Clancy T, Hill J, Evans D (2011) Genotype-phenotype correlation in colorectal polyposis. *Clin Genet*.
- Nichols KE, Houseknecht MD, Godmilow L, Bunin G, Shields C, Meadows A, Ganguly A (2005) Sensitive multistep clinical molecular screening of 180 unrelated individuals with retinoblastoma detects 36 novel mutations in the RB1 gene. *Hum Mutat* 25: 566-74.
- Nieuwenhuis MH, Vogt S, Jones N, Nielsen M, Hes FJ, Sampson JR, Aretz S, Vasen HF (2012) Evidence for accelerated colorectal adenoma--carcinoma progression in MUTYH-associated polyposis? *Gut* 61: 734-8.
- Oikawa T, Nakamura A, Onishi N, Yamada T, Matsuo K, Saya H (2013) Acquired expression of NFATc1 downregulates E-cadherin and promotes cancer cell invasion. *Cancer Res* 73: 5100-9.
- Orozco LD, Cokus SJ, Ghazalpour A, Ingram-Drake L, Wang S, van Nas A, Che N, Araujo JA, Pellegrini M, Lusk AJ (2009) Copy number variation influences gene expression and metabolic traits in mice. *Hum Mol Genet* 18: 4118-29.
- Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z (2013) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*.
- Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I, Sherborne A, Chubb D, Carvajal-Carmona LG, Ma Y, Kaur K, Dobbins S, Barclay E, Gorman M, Martin L, Kovac MB, Humphray S, Lucassen A, Holmes CC, Bentley D, Donnelly P, Taylor J, Petridis C, Roylance R, Sawyer EJ, Kerr DJ, Clark S, Grimes J, Kearsey SE, Thomas HJ, McVean G, Houlston RS, Tomlinson I (2013) Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 45: 136-44.
- Pasternack SM, von Kugelgen I, Al Aboud K, Lee YA, Ruschendorf F, Voss K, Hillmer AM, Molderings GJ, Franz T, Ramirez A, Nurnberg P, Nothen MM, Betz RC (2008) G protein-coupled receptor P2Y5 and its ligand LPA are involved in maintenance of human hair growth. *Nat Genet* 40: 329-34.

- Paulson TG, Galipeau PC, Reid BJ (1999) Loss of heterozygosity analysis using whole genome amplification, cell sorting, and fluorescence-based PCR. *Genome Res* 9: 482-91.
- Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 16: 1136-48.
- Pendas-Franco N, Aguilera O, Pereira F, Gonzalez-Sancho JM, Munoz A (2008) Vitamin D and Wnt/beta-catenin pathway in colon cancer: role and regulation of DICKKOPF genes. *Anticancer Res* 28: 2613-23.
- Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenge L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* 82: 685-95.
- Petersen GM, Slack J, Nakamura Y (1991) Screening guidelines and premorbid diagnosis of familial adenomatous polyposis using linkage. *Gastroenterology* 100: 1658-64.
- Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9: e1003709.
- Pinkel D, Albertson DG (2005) Comparative genomic hybridization. *Annu Rev Genomics Hum Genet* 6: 331-54.
- Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 29: 512-20.
- Pinto D, Marshall C, Feuk L, Scherer SW (2007) Copy-number variation in control population cohorts. *Hum Mol Genet* 16 Spec No. 2: R168-73.
- Prenen H, Vecchione L, Van Cutsem E (2013) Role of targeted agents in metastatic colorectal cancer. *Target Oncol* 8: 83-96.
- Priebe L, Degenhardt FA, Herms S, Haenisch B, Mattheisen M, Nieratschker V, Weingarten M, Witt S, Breuer R, Paul T, Alblas M, Moebus S, Lathrop M, Leboyer M, Schreiber S, Grigoriou-Serbanescu M, Maier W, Propping P, Rietschel M, Nothen MM, Cichon S, Muhleisen TW (2012) Genome-wide survey implicates the influence of copy number variants (CNVs) in the development of early-onset bipolar disorder. *Mol Psychiatry* 17: 421-32.
- Priolli DG, Canello TP, Lopes CO, Valdivia JC, Martinez NP, Acari DP, Cardinalli IA, Ribeiro ML (2013) Oxidative DNA damage and beta-catenin expression in colorectal cancer evolution. *Int J Colorectal Dis* 28: 713-22.
- Pulst SM (1999) Genetic linkage analysis. *Arch Neurol* 56: 667-72.
- Pylayeva-Gupta Y, Grabocka E, Bar-Sagi D (2011) RAS oncogenes: weaving a tumorigenic web. *Nat Rev Cancer* 11: 761-74.
- Pylkas K, Vuorela M, Otsukka M, Kallioniemi A, Jukkola-Vuorinen A, Winqvist R (2012) Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. *PLoS Genet* 8: e1002734.
- Ramanan VK, Shen L, Moore JH, Saykin AJ (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet* 28: 323-32.
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Armengol L, Conrad DF, Estivill X, Tyler-

- Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME (2006) Global variation in copy number in the human genome. *Nature* 444: 444-54.
- Renner C, Pfitzenmeier JP, Gerlach K, Held G, Ohnesorge S, Sahin U, Bauer S, Pfreundschuh M (1997) RP1, a new member of the adenomatous polyposis coli-binding EB1-like gene family, is differentially expressed in activated T cells. *J Immunol* 159: 1276-83.
- Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D (2012) Ancient DNA studies: new perspectives on old samples. *Genet Sel Evol* 44: 21.
- Rodriguez-Pineiro AM, de la Cadena MP, Lopez-Saco A, Rodriguez-Berrocal FJ (2006) Differential expression of serum clusterin isoforms in colorectal cancer. *Mol Cell Proteomics* 5: 1647-57.
- Rokman A, Baffoe-Bonnie AB, Gillanders E, Fredriksson H, Autio V, Ikonen T, Gibbs KD, Jr., Jones M, Gildea D, Freas-Lutz D, Markey C, Matikainen MP, Koivisto PA, Tammela TL, Kallioniemi OP, Trent J, Bailey-Wilson JE, Schleutker J (2005) Hereditary prostate cancer in Finland: fine-mapping validates 3p26 as a major predisposition locus. *Hum Genet* 116: 43-50.
- Ronsch K, Jager M, Schopflin A, Danciu M, Lassmann S, Hecht A (2011) Class I and III HDACs and loss of active chromatin features contribute to epigenetic silencing of CDX1 and EPHB tumor suppressor genes in colorectal cancer. *Epigenetics* 6: 610-22.
- Rousseau-Merck MF, Huebner K, Berger R, Thiesen HJ (1991) Chromosomal localization of two human zinc finger protein genes, ZNF24 (KOX17) and ZNF29 (KOX26), to 18q12 and 17p13-p12, respectively. *Genomics* 9: 154-61.
- Rutzky LP, Siciliano MJ (1982) Various isozyme gene expression patterns among human colorectal adenocarcinoma cell lines and tissues. *J Natl Cancer Inst* 68: 81-90.
- Sadeghi A, Frohlich H (2013) Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics* 14: 144.
- Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94: 441-8.
- Sato H, Suzuki H, Toyota M, Nojima M, Maruyama R, Sasaki S, Takagi H, Sogabe Y, Sasaki Y, Idogawa M, Sonoda T, Mori M, Imai K, Tokino T, Shinomura Y (2007) Frequent epigenetic inactivation of DICKKOPF family genes in human gastrointestinal tumors. *Carcinogenesis* 28: 2459-66.
- Segditsas S, Rowan AJ, Howarth K, Jones A, Leedham S, Wright NA, Gorman P, Chambers W, Domingo E, Roylance RR, Sawyer EJ, Sieber OM, Tomlinson IP (2009) APC and the three-hit hypothesis. *Oncogene* 28: 146-55.
- Seiden-Long I, Navab R, Shih W, Li M, Chow J, Zhu CQ, Radulovich N, Saucier C, Tsao MS (2008) Gab1 but not Grb2 mediates tumor progression in Met overexpressing colorectal cancer cells. *Carcinogenesis* 29: 647-55.
- Seko A, Nagata K, Yonezawa S, Yamashita K (2002) Down-regulation of Gal 3-O-sulfotransferase-2 (Gal3ST-2) expression in human colonic non-mucinous adenocarcinoma. *Jpn J Cancer Res* 93: 507-15.
- Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'Arcy M, Frackelton EC, Geiger EA, Haldeman-Englert C, Imielinski M, Kim CE, Medne L, Annaiah K, Bradfield JP, Dabaghyan E, Eckert A, Onyiah CC, Ostapenko S, Otieno FG, Santa E, Shaner JL, Skraban R, Smith RM, Elia J, Goldmuntz E, Spinner NB, Zackai EH, Chiavacci RM, Grundmeier R, Rappaport EF, Grant SF, White PS, Hakonarson H (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* 19: 1682-90.
- Shapiro MB, Senapathy P (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* 15: 7155-74.



- Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 38: 1038-42.
- Shimao Y, Nabeshima K, Inoue T, Koono M (2002) Complex formation of IQGAP1 with E-cadherin/catenin during cohort migration of carcinoma cells. Its possible association with localized release from cell-cell adhesion. *Virchows Arch* 441: 124-32.
- Shlien A, Malkin D (2009) Copy number variations and cancer. *Genome Med* 1: 62.
- Shlien A, Malkin D (2010) Copy number variations and cancer susceptibility. *Curr Opin Oncol* 22: 55-63.
- Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, Novokmet A, Nanda S, Druker H, Scherer SW, Malkin D (2008) Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci U S A* 105: 11264-9.
- Shoji Y, Takahashi M, Kitamura T, Watanabe K, Kawamori T, Maruyama T, Sugimoto Y, Negishi M, Narumiya S, Sugimura T, Wakabayashi K (2004) Downregulation of prostaglandin E receptor subtype EP3 during colon cancer development. *Gut* 53: 1151-8.
- Sjoberg T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314: 268-74.
- Smith CG, Naven M, Harris R, Colley J, West H, Li N, Liu Y, Adams R, Maughan TS, Nichols L, Kaplan R, Wagner MJ, McLeod HL, Cheadle JP (2013) Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer. *Hum Mutat* 34: 1026-34.
- Stadler ZK, Esposito D, Shah S, Vijai J, Yamrom B, Levy D, Lee YH, Kendall J, Leotta A, Ronemus M, Hansen N, Sarrel K, Rau-Murthy R, Schrader K, Kauff N, Klein RJ, Lipkin SM, Murali R, Robson M, Sheinfeld J, Feldman D, Bosl G, Norton L, Wigler M, Offit K (2012) Rare de novo germline copy-number variation in testicular cancer. *Am J Hum Genet* 91: 379-83.
- Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med* 61: 437-55.
- Strachan T, Read AP (1999) Human molecular genetics, 2nd edn. Wiley, New York
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315: 848-53.
- Suceveanu AI, Suceveanu A, Voinea F, Mazilu L, Mixici F, Adam T (2009) Introduction of cytogenetic tests in colorectal cancer screening. *J Gastrointest Liver Dis* 18: 33-8.
- Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 3: 391-7.
- Takafuji V, Lublin D, Lynch K, Roche JK (2001) Mucosal prostanoid receptors and synthesis in familial adenomatous polyposis. *Histochem Cell Biol* 116: 171-81.
- Takei Y, Assenberg M, Tsujimoto G, Laskey R (2002) The MCM3 acetylase MCM3AP inhibits initiation, but not elongation, of DNA replication via interaction with MCM3. *J Biol Chem* 277: 43121-5.
- Tanikawa C, Furukawa Y, Yoshida N, Arakawa H, Nakamura Y, Matsuda K (2009) XEDAR as a putative colorectal tumor suppressor that mediates p53-regulated anoikis pathway. *Oncogene* 28: 3081-92.

- Tanikawa C, Ri C, Kumar V, Nakamura Y, Matsuda K (2010) Crosstalk of EDA-A2/XEDAR in the p53 signaling pathway. *Mol Cancer Res* 8: 855-63.
- Thean LF, Loi C, Ho KS, Koh PK, Eu KW, Cheah PY (2010) Genome-wide scan identifies a copy number variable region at 3q26 that regulates PPM1L in APC mutation-negative familial colorectal cancer patients. *Genes Chromosomes Cancer* 49: 99-106.
- Therkildsen C, Jonsson G, Dominguez-Valentin M, Nissen A, Rambech E, Halvarsson B, Bernstein I, Borg K, Nilbert M (2013) Gain of chromosomal region 20q and loss of 18 discriminates between Lynch syndrome and familial colorectal cancer. *Eur J Cancer* 49: 1226-35.
- Thirlwell C, Howarth KM, Segditsas S, Guerra G, Thomas HJ, Phillips RK, Talbot IC, Gorman M, Novelli MR, Sieber OM, Tomlinson IP (2007) Investigation of pathogenic mechanisms in multiple colorectal adenoma patients without germline APC or MYH/MUTYH mutations. *Br J Cancer* 96: 1729-34.
- Tuffery-Giraud S, Saquet C, Chambert S, Claustres M (2003) Pseudoexon activation in the DMD gene as a novel mechanism for Becker muscular dystrophy. *Hum Mutat* 21: 608-14.
- Tuohy TM, Done MW, Lewandowski MS, Shires PM, Saraiya DS, Huang SC, Neklason DW, Burt RW (2010) Large intron 14 rearrangement in APC results in splice defect and attenuated FAP. *Hum Genet* 127: 359-69.
- Tuupanen S, Yan J, Turunen M, Gylfe AE, Kaasinen E, Li L, Eng C, Culver DA, Kalady MF, Pennison MJ, Pasche B, Manne U, de la Chapelle A, Hampel H, Henderson BE, Le Marchand L, Hautaniemi S, Askhtorab H, Smoot D, Sandler RS, Keku T, Kupfer SS, Ellis NA, Haiman CA, Taipale J, Aaltonen LA (2012) Characterization of the colorectal cancer-associated enhancer MYC-335 at 8q24: the role of rs67491583. *Cancer Genet* 205: 25-33.
- Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37: 727-32.
- Valsesia A, Mace A, Jacquemont S, Beckmann JS, Kotalik Z (2013) The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. *Front Genet* 4: 92.
- van der Meulen-de Jong AE, Morreau H, Becx MC, Crobach LF, van Haastert M, ten Hove WR, Kleibeuker JH, Meijssen MA, Nagengast FM, Rijk MC, Salemans JM, Stronkhorst A, Tuynman HA, Vecht J, Verhulst ML, de Vos tot Nederveen Cappel WH, Walinga H, Weinhardt OK, Westerveld BD, Witte AM, Wolters HJ, Vasen HF (2011) High detection rate of adenomas in familial colorectal cancer. *Gut* 60: 73-6.
- van Hattem WA, Brosens LA, de Leng WW, Morsink FH, Lens S, Carvalho R, Giardiello FM, Offerhaus GJ (2008) Large genomic deletions of SMAD4, BMPR1A and PTEN in juvenile polyposis. *Gut* 57: 623-7.
- Vasanthakumar A, Lepore JB, Zegarek MH, Kocherginsky M, Singh M, Davis EM, Link PA, Anastasi J, Le Beau MM, Karpf AR, Godley LA (2013) Dnmt3b is a haploinsufficient tumor suppressor gene in Myc-induced lymphomagenesis. *Blood* 121: 2059-63.
- Vasen HF, Mecklin JP, Khan PM, Lynch HT (1991) The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Dis Colon Rectum* 34: 424-5.
- Vasen HF, Watson P, Mecklin JP, Lynch HT (1999) New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology* 116: 1453-6.
- Venkatachalam R, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A, Kuiper RP (2008) Predisposition to colorectal cancer: exploiting copy number variation to identify novel predisposing genes and mechanisms. *Cytogenet Genome Res* 123: 188-94.

- Venkatachalam R, Verwiel ET, Kamping EJ, Hoenselaar E, Gorgens H, Schackert HK, van Krieken JH, Ligtenberg MJ, Hoogerbrugge N, van Kessel AG, Kuiper RP (2011) Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int J Cancer* 129: 1635-42.
- Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL (1988) Genetic alterations during colorectal-tumor development. *N Engl J Med* 319: 525-32.
- Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW (2013) Cancer genome landscapes. *Science* 339: 1546-58.
- Wang K, Bucan M (2008) Copy Number Variation Detection via High-Density SNP Genotyping. *CSH Protoc* 2008: pdb top46.
- Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17: 1665-74.
- Wang Q, Zhao ZB, Wang G, Hui Z, Wang MH, Pan JF, Zheng H (2013) High expression of KIF26B in breast cancer associates with poor prognosis. *PLoS One* 8: e61640.
- Watanabe T, Kobunai T, Yamamoto Y, Matsuda K, Ishihara S, Nozawa K, Iinuma H, Ikeuchi H, Eshima K (2011) Differential gene expression signatures between colorectal cancers with and without KRAS mutations: crosstalk between the KRAS pathway and other signalling pathways. *Eur J Cancer* 47: 1946-54.
- Wei Q, Zhou W, Wang W, Gao B, Wang L, Cao J, Liu ZP (2010) Tumor-suppressive functions of leucine zipper transcription factor-like 1. *Cancer Res* 70: 2942-50.
- Xi Y, Formentini A, Nakajima G, Kornmann M, Ju J (2008) Validation of biomarkers associated with 5-fluorouracil and thymidylate synthase in colorectal cancer. *Oncol Rep* 19: 257-62.
- Xie D, Sham JS, Zeng WF, Che LH, Zhang M, Wu HX, Lin HL, Wen JM, Lau SH, Hu L, Guan XY (2005) Oncogenic role of clusterin overexpression in multistage colorectal tumorigenesis and progression. *World J Gastroenterol* 11: 3285-9.
- Xu Y, Peng B, Fu Y, Amos CI (2011) Genome-wide algorithm for detecting CNV associations with diseases. *BMC Bioinformatics* 12: 331.
- Yam YY, Hoh BP, Othman NH, Hassan S, Yahya MM, Zakaria Z, Ankathil R (2013) Somatic copy-neutral loss of heterozygosity and copy number abnormalities in Malaysian sporadic colorectal carcinoma patients. *Genet Mol Res* 12: 319-27.
- Yang R, Chen B, Pfitze K, Buch S, Steinke V, Holinski-Feder E, Stocker S, von Schonfels W, Becker T, Schackert HK, Royer-Pokora B, Kloor M, Schmiegeler WH, Buttner R, Engel C, Lascorz Puertolas J, Forsti A, Kunkel N, Bugert P, Schreiber S, Krawczak M, Schafmayer C, Propping P, Hampe J, Hemminki K, Burwinkel B (2013) Genome-wide analysis associates familial colorectal cancer with increases in copy number variations and a rare structural variation at 12p12.3. *Carcinogenesis*.
- Yang XR, Brown K, Landi MT, Ghiorzo P, Badenas C, Xu M, Hayward NK, Calista D, Landi G, Bruno W, Bianchi-Scarra G, Aguilera P, Puig S, Goldstein AM, Tucker MA (2012) Duplication of CXC chemokine genes on chromosome 4q13 in a melanoma-prone family. *Pigment Cell Melanoma Res* 25: 243-7.
- Yau C, Holmes CC (2008) CNV discovery using SNP genotyping arrays. *Cytogenet Genome Res* 123: 307-12.
- Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11: 377-94.
- Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 10: 451-81.

- 
- Zhang K, Nowak I, Rushlow D, Gallie BL, Lohmann DR (2008) Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression. *Hum Mutat* 29: 475-84.
- Zhang L, Yang W, Ying D, Cherny SS, Hildebrandt F, Sham PC, Lau YL (2011) Homozygosity mapping on a single patient: identification of homozygous regions of recent common ancestry by using population data. *Hum Mutat* 32: 345-53.
- Zhao M, Wang Q, Jia P, Zhao Z (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14 Suppl 11: S1.
- Zhu M, Zhao S (2007) Candidate gene identification approach: progress and challenges. *Int J Biol Sci* 3: 420-7.
- Zuo ZG, Yu ZQ, Gao XH, Wang H, Zhang C, Liu QZ, Han YF, Chen LP, Zhang XQ, Fu CG (2013) [Association of epidermal growth factor receptor expression and its downstream gene mutation status with radiosensitivity of colorectal carcinoma cell lines in vitro]. *Zhonghua Wei Chang Wai Ke Za Zhi* 16: 753-8.

# List of publications

- Spier S, Holzapfel S, Altmüller J, Zhao B, **Horpaopan S**, Vogt S, Chen S, Morak M, Raeder S, Kayser K, Stienen D, Adam R, Uhlhaas S, Sengteller M, Nürnberg P, Holinski-Feder E, Nöthen MM, Lifton RP, Thiele H, Hoffmann P, Steinke-Lange V, Aretz S. Systematic screening of eight polymerase genes identified germline *POLE* mutations as relevant cause of unexplained familial colorectal adenomas and carcinomas. *Int J Cancer*. 2014 Dec 20. doi: 10.1002/ijc.29396.
- Horpaopan S**, Spier S, Zink AM, Altmüller J, Holzapfel S, Laner A, Vogt S, Uhlhaas S, Heilmann S, Stienen D, Pasternack S, Keppler K, Adam R, Kayser K, Moebus S, Draaken M, Degenhardt F, Engels H, Hofmann A, Nöthen MM, Steinke-Lange V, Herms S, Holinski-Feder E, Fröhlich H, Thiele H, Hoffmann P, Aretz S. Genome-wide CNV analysis in 221 unrelated patients and targeted high-throughput sequencing reveal novel causative candidate genes for colorectal adenomatous polyposis. *Int J Cancer*. 2015 Mar; 136(6): E578-89.
- Aretz S, Tricarico R, Papi L, Spier I, Pin E, **Horpaopan S**, Cordisco EL, Pedroni M, Stienen D, Gentile A, Panza A, Piepoli A, de Leon MP, Friedl W, Viel A, Genuardi M. *MUTYH*-associated polyposis (MAP): evidence for the origin of the common European mutations p.Tyr179Cys and p.Gly396Asp by founder events. *Eur J Hum Genet*. 2013 Jan 30.
- Spier I, **Horpaopan S**, Vogt S, Uhlhaas S, Morak M, Stienen D, Draaken M, Ludwig M, Holinski-Feder E, Nöthen MM, Hoffmann P, Aretz S. Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. *Hum Mutat*. 2012 Jul; 33(7): 1045-50.
- Pangjaidee N, **Horpaopan S**, Puwachiranon N, Puaninta C, Tanpaiboon P, Chanprapaph P, Vutyananich T, Mevatee U. Identification of Marker Chromosome by Micro-FISH Technique. Abstracts Supplement 27th Annual Scientific Meeting.

# Appendices

**Table A1.** Forward and Reverse primers of 5 overlapping fragments of *APC* gene

Frag ment	Size (bp)	For primer (5'-3')	Location	Rev primer (5'-3')	Location
1	575	TCTGTACCACCCTCAGTTCTG	5'UTR	AGAGAGGTCATTGCTTCTTGC	Exon 3
2	615	TCTGGAGAGTGCAGTCCTGTT	Exon 3	CAAGTCATCTGGGAACCAAGG	Exon 8
3	665	GTGGGAGAAATCAACATGGCA	Eon 7	GTGGACTGTGAAATGTATGGG	Exon 11
4	505	CAGATCTGTCCTGCTGTGTGT	Exon 10	ATGTGCTGTAGATGGTGCAC	Exon 14
5	564	GACGTTGCGAGAAGTTGGAAG	Exon 13	ATGTCTCCTGGCTCAAGCTTG	Exon 15A

**Table A2.** Primers used for Sanger sequencing of the inserted/deleted region on the genomic level

Location	For primer (5'-3')	Rev primer (5'-3')
Intron 4 <sup>#</sup>	TTTATGTTGGGAAGCCAAGC	CTTTTATTGCCTTTTGGGCA
Intron 10	CAGCAGTGCACCTCCATTTTT	CCTATCCCCTCATCGTTTCA

<sup>#</sup> only the reverse sequence was used

**Table A3.** Primers used for the amplication of microsatellite markers and the expected size of the PCR products

Marker	Forward primer	Reverse primer	Size (bp)
D5S134	ACATCTCCAATATACCCCCCTCTCTTTTC	TCCTCTGTGGTTGGTGAAATTGCACC	167-183
D5S492	TTTCCCCAATACAACGTGA	AACCAGCAAACCTCAGAAGTG	248-260
D5S1965	TGTCCCCTTGATAAAAAATTACTGCG	GTGTCTGGGATTTCTACGCAATG	230-252
D5S346	ACTCACTCTAGTGATAAATCGGG	AGCAGATAAGACAGTATTACTAGTT	96-110
D5S656	GCTAAGAAAATACGACAACATAATG	CATAATAAACTGATGTTGACACAC	185-203
D5S2001	GCCAAGATGGTCTCGATCTC	TCTGAACAGGTGATGGCAAC	255-273
D5S421	TGGAAATAGAATCCAGGCTT	TCTATCGTTAACTTTATTGATTGAG	152-170

**Table A4.** Primer sequences used for validating 46 deletion CNVs by qPCR with SYBR Green I

Deletion CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr1:45592076-45700435	108360	18	57.1361	1st 2nd 3rd	ACCCCTTCAATGGTTCCCTTC CAGAAGAGGGAAGGCTGTTG CGAAGATGGCCAACTCTTC	AGAGCCAGAGGAAAAGGAC GAGCTGGGTTCAAATCTCA ACTATGGATCGGAGCAAACG
chr1:71094328-71116548	22221	52	213.048	1st 2nd 3rd	ATCCCAGGGTTAGGCAGATT AACGGAAGGCAAAACAGAAG GCGTATGGCCAGCAAGTTAT	TTCAACAGTGTGAAAGCAGGA CCAGGACTTGGTGCAGTTCT AGAAAGCCCGGAAGTAAAGC
chr1:92580922-92800950	220029	48	207.295	1st 2nd 3rd	GCGGAATGGACTTTAATTGC TAGCATTCTACGCCAAGG CTGGGATTACACGCAGGAAT	TCCAAGGCAGTTTACAAAATCA CGGAGTTTGAGGACTTCTGG CAAAATGTCCCTTGCAACCT
chr1:246404335-246638006	233672	132	198.607	1st 2nd 3rd	AACCGGAATGGAAGTTGAGA AGACGATTGGAGAGGGGAAC GCCTCCCTCTTTAACCTA	CTCTTCAGTGCTGGGAAGGT CTGGAACAGCTCACCAACA TGAGCCTCTCGGGTTTCTTA
chr2:137473187-137569168	95982	33	122.085	1st 2nd 3rd	CCCAGTGTGGTGATGACTGA TGTGTGCTTGTTCTTACGC GCATGGGGCTTTCTTTATGA	GGTGGCCAGAAATTGTGATT ACCACAGTTCGGTTCAGCTT ACCCAGCAAGTTTGTGGAAC
chr2:137484018-137615223	131206	51	193.92	1st 2nd 3rd	CCCAGTGTGGTGATGACTGA TGTGTGCTTGTTCTTACGC GCATGGGGCTTTCTTTATGA	GGTGGCCAGAAATTGTGATT ACCACAGTTCGGTTCAGCTT ACCCAGCAAGTTTGTGGAAC
chr2:158813339-158905017	91679	51	196.779	1st 2nd 3rd	TCCTGTCATCCCATCACGTA CCAGTTCTGCCATTCAACT GAAGGAAAATGGGCTGTGAA	GGGCAGGAAAGTTTTAAGCA AGAGGTGGGGAGTCAGGTTT AAGCTGGCTGAGATTGTGGT
chr3:260287-291417	31131	32	82.4472	1st 2nd 3rd	GAAGTGTGCTCCATGCCCTTG ATGTTCCGGTGGATTGGATGT GACCCCAAGTTCAAGCTGAG	GCTTCTGAATGTCCCAGAG TGATGCAGTTGGCCATTAAA AGGAAAGGCCTCCCTGATAA
chr3:41692118-41708287#	16170	20	28.2084	1st 2nd 3rd	TGGGTAGGTCATGGGAAAC AGAGCAACCGCAAGCTAAAG AACCAGGAACATCAGGCAAC	AATTGGAGGGAGCATGTGT CACCAGCAAGAATTGCATCA CGAGGAAGCTTCTGATCACC
chr3:45839485-45852716	13232	8	20.4535	1st 2nd 3rd	ATGGTTTCTGAAGGCCACAC TAGCTGGTCCTGCAATCCT TCCCAGCTCATAGTTGTC	TTGAGGGCTTCATTCTCACC GGCCTCACCTCAGTCAGAAG ATGCCAGAAATCAGGAAGGA
chr3:176596105-176753724	157620	72	336.103	1st 2nd 3rd	GACAAAGCCGATCAAACCTC CCATCATCAGCTGCACACA TTGGGTTTGGATTGGTGTTT	GCTCTAGAGAGCTACACGCTTCAT TCCTCCCCAAGAACAGAAAA GTTGTGTTGAAGGGCTCAT
chr3:194612501-194625470#	12970	19	40.0346	1st 2nd 3rd	GGACATGTCTGTTCCCTCT CCTCTGAACCAGGATGAAGC ACCTGATGGCTTCTCCACTG	CTGGTCTTGGGAACAATCAA CCTAAGCTGGTGCTTTTCAG TGTCTTGTAGTGGGGACT
chr4:83794836-83805695	10860	9	32.9419	1st 2nd 3rd	ACAGGGATTGGCAAAATGAA ACCAAGCCTCGCTGTTGTAT TCTGGCACATCAGGCTCTAA	TGTAGGCCAGAGGGTAATG AGTCACCTTCCCGGATTAC GGGGTTGGGTCAATTAATCA
chr4:91075998-91166816	90819	20	92.0585	1st 2nd 3rd	TTTGAGACTCGTTGCAAGA AAGGAACAATTCCAGCCAAG TTGCAGTGGAGGGGACTAAA	TTCAACGCCTTAAATCTGG AGCAGATGTGCAAGCAAGA TCCAGAAGTGGGAAAGGTA
chr4:162981340-163113553	132214	38	97.0885	1st 2nd	CAAAGAGGGTGGTTAGCAA TCTCCGTCAGATCCACACAC	TGTACAACAGCTCCCATCA GCCCTTTTGGATCTGTGAA

Deletion CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr4:162981340-163113553	132214	38	97.0885	3rd	AATCTCCAGCACCTGACCAC	CTGTCTTGGGCTCTCTCCAA
chr4:172930640-172979152	48513	12	21.5379	1st	GCCTCAGGAAGTGCATTGTA	GGGCTACATGTGCTTGAGGT
				2nd	CCAAGCCTTTGCACTCTTTC	GGCGTAGAGCGACATCTTTC
				3rd	GGGGAGAGGAAAGGAATTTG	TAAAGCCACCGTGAACAACA
chr5:5229672-5280314	50643	46	182.73	1st	TTCAAAAGAACCCAGGCATC	GTCTGCCCTTTTGTATGTGGT
				2nd	ACATCCCCAGGTAGGGGTAG	GAGAGATCCAGGCGTTTGAG
				3rd	TGCTGTGAAAGCCACAGAAC	GTCCACATTGCTGCAACATC
chr5:58578074-58734281	156208	72	326.102	1st	TCACAAAGCCTCGACATCAC	GCCAAGATCCCAGTCTACAA
				2nd	GGGCCTAAGGCAGTTAAACA	TTTTGTTGGGATCCTCTGG
chr5:112323863-112351637	27775	11	21.2786	1st	CTTGTCATTGCTGCTTGAA	GCTGTGCTCCCTAGTGAAGG
				2nd	TGCTGTGGGTCCTCATCAT	GTCCGCGAAACTGAGAAAAC
				3rd	TGGTCCCTTTTCTCAGTTGG	CTTTCCTCCTGCGGAATATG
chr5:119920805-120247470	326666	107	609.173	1st	AGTGTGCTTGGTGTGTTCCA	AGCCAGGTCTTCACAACAGC
				2nd	TGACTGACAGCTCCAAAACG	AGGACCGTGAGGATAGCAGA
				3rd	TTGGGGCAGAGAGAGCTTAG	TTGCCTTTTCATTTCAGATG
chr9:19101637-19117877	16241	12	48.4789	1st	GCATTGCGGAACACTGAGTA	ACTCAGCAGCTCCAGACCAC
				2nd	GATCCAGGTTGGGAAAACAA	ACTCAGGTGAGCTTGGGACA
				3rd	ACCAGTCCCTTGACTTGCAT	GAGATGGCAGAGAACGGTGT
chr9:20856974-20934558	77585	30	122.794	1st	CCCTCTGGTCAGAATCAAA	GGCGGAATCAGCCATACCTTA
				2nd	GGGCATGACTTTCTTCAAGC	CCTGAACCTCCACGTTTTTA
				3rd	GGCTTTGTCTCTCCTTGTGC	GATTTGCACGGCTTGAGACT
chr9:28465532-28493578	28047	16	61.4615	1st	AATTGTTGGGCTCTTCATGC	CCTTCTGAAGCAAGCGAGT
				2nd	GCAGTTCCAGAGGAAGATGG	GCTTTCTTCCATCCAAGCAG
				3rd	TGAAAGGAAGATGGCAAAGC	GAGCAAAGAAAGCCAACAGG
chr9:94670184-94740522	70339	24	60.0938	1st	ACCCTAATGGTCCTGGCTCT	ACAGGTGTGCCTAGGATTGG
				2nd	TCTCCATAGCTCTCCCCAGA	TTTTGGTCTTTGCCTCCAG
				3rd	CACAATTCAAGAGCCCAGGT	CAGACTGTGTCCTCCCTCCATT
chr10:73839878-73855644	15767	10	26.3311	1st	TGGATGATTCCCTCTTCCTG	AAGAATGGTGAGAGGCATGG
				2nd	GTGCTTCTTGAGCTGCCTCT	TCTGTTTCCTCCAGTTTG
chr10:76856600-77074588	217989	50	117.305	1st	TTCGCAAAGAGGCTTGATCT	AAAGGACAACCCATGTGGAG
				2nd	AATTTTGTCCCATCCATCCA	GCACCTCCAAACTCAGGTTT
				3rd	CCCAGGGCTTGATTGTGT	GGCCATCTTGGTATTGTCAT
chr11:44042462-44061557	19096	21	34.4747	1st	GATTGAGCTGAGGAGGCTA	CTGGATGCTGCCTATTGTT
				2nd	CCTGGAAGTGGTTCATCTTCA	GCCACAGAGCTCACTTACCC
				3rd	CAGAGGACCCATGTGATGTG	CTGCAGGAAGAAGCACCTTA
chr11:8116419-8185496	69078	52	217.569	1st	CCCAAACCGTAGATTGGAA	TCGTTTCCAGAGGTCTCACC
				2nd	TGGCCTGCTCTTACCTTCAG	CTCCACAGTGACAGAGTCG
				3rd	TGCCCCACATACCTTTTCCTC	CTCTCGGGCTCAGATCAAAC
chr11:102695264-102848464	153201	64	209.884	1st	TTCCAGAGGACGACAAACCT	CTCCTTCAGGCTCACTAGCAA
				2nd	GAAAGTGCCTCCTCCTAACG	TTGCTGAGAGCAGCAAGAGA
				3rd	ATGGCAGATCAGTGGCTTGT	TGGAATCCAGCCCATAAAAC
chr12:20897382-21303121	405740	353	1606.22	1st	ATGTTCTTGGCAGCCCTGT	CCAATTTCAAAGCTTCCATCA
				2nd	AAGGCATCGACAATGAAAG	GCAATGTTAGTTGGCAGCAG
				3rd	GACGGAAGCTTTGAAATTGG	TTATTGCCAATTGCCTGTG
				4th	TCAATGTGGAATATCATGCAG	CAGTTGTTGGTGGACCACTT



Deletion CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr12:20908555-20927800	19246	23	90.009	1st 2nd 3rd	AACACCATGATCCCTTTGGA CCACTGGGGCCCTTAAGTAA TTGGGTCATGGTGAAGTCTG	TGCACAGAACATCAGCGTTA TTTGCTTTCGCAGATTAGAGG TGTGGTACCTCCTGTTGCAG
chr12:53445641-54110181	664541	256	1432.03	1st 2nd 3rd	CAAGGACTGAGGGTCTTCCA ATTTCTGCTGGGTGATGGAC TGGCTTGTAAATCAGCTCCTG	AGCAGCTGTTGTCGGATTCT CACCAATACGGAACCCATC TGTGTCTGTGCAAGCAATCA
chr12:80663903-80677221#	13319	16	27.5026	1st 2nd 3rd	ACCAGGAAGCATACCTGTGG AAGGGCTGTCAGGGCATAAT ATTTCCGGTCTTGTCTCTT	TCTAGACACCCCTCGGGAGA AATTTGGGATGTGCTGGTGT TCTGAGCTGCCCTTCTCCTC
chr13:84714710-85327835	613126	219	882.91	1st 2nd 3rd	CTGGCAATTCTCAGTGGTGA TGTCTAATGGTGCCCTGCAG GTTGCGCTCCTGACAATGTA	GTGGCACAAGGAAAGTGTC AATCCATCGTGCCATAGTCC ACTACCCACCAAGCACCAG
chr14:87469000-87489111	20112	22	97.8503	1st 2nd 3rd	TGGCTTCCACAAGAAAGTTG TCCATTTGCAAAAATCCAGA TGACATTTCTGTGCCCTTTT	TGGCATGCTGAATGACAAGT TTTTGCAGGACCAATCTGAC CTTGAAAAACATCCGAAAGA
chr15:40160584-40195058	34475	28	92.5148	1st 2nd 3rd	AAGCCTAGGATTGCAACGA TCACACTTCCAGAATGCAC GGAGTTCCAGGAATGGATCA	TGCTGAAGGGTCTCTATGAG CTAGCATGGAGAGCCTGTCA CTTGTTTGTGGGTGTTCTG
chr16:3072198-3088992	16795	10	21.4463	1st 2nd 3rd	ATGGACCTACCCACATTCCA AGCCACTTTGACCTCCAAAA AGCCTCTCCCTGGTAACGAT	TTCCGGCTCTGTAATCAAGG GCAGAGTCCAGGCAGATAG ATCCCCCATCTGTACCACA
chr17:5423583-5690836	267254	190	543.478	1st 2nd	CCTGACGTTTCATCCAGAGG CCAGGTATGGAGGGCTAGGT	GTCTCAAGCCAGGAGTCAC AGCTTCTGCTCGCCAATAAA
chr17:29847372-30198038	350667	192	378.297	1st 2nd 3rd	CAGTTGCCTTGCTTCTAGG GTCAGAGGAGCTGGCCTATG AGCCATATGGAGAGGTGGTG	CCCAGCTTAGATTCCAGCAC CGCATGTAATTGTGAGGTC GGGTCCAAGAGACTGCAGAG
chr18:14712857-14748540	35684	29	38.9638	1st 2nd 3rd	CCTCCTGTTTTGCTGAGAC AGCGAACGGGTCTACACTGA TTGCCATATGTGAAGGCTCA	GGCCTCTTCCCAGAGTAACC ACGGGCTTCTTCCCTACTGT TAACTCCACAAGCAGCAGCA
chr19:15931793-15964505	32713	12	23.4107	1st 2nd 3rd	AGGACTGTGGCTGATCTGGT AACTGCACACCACAACCAAA CTCTGGGCTGCTTCAGTACC	TCACGAATGGGACTGGTGTA TTCCGCATCCATTGTTATGA CAGGGATGTGGGAGATGAAC
chr19:62581812-62597371	15560	10	39.1491	1st 2nd 3rd	CACTGACCATTCTCTGTGC CGAAGACCTTCAAGGGAAT ATTGGTGGCTCCACAGACAT	TGATAGTTCTGTGGCTTTGG ATGAGCCACTGTACCCACAA CGAAACCACCTCACACTTCA
chr19:62989408-63011738	22331	9	23.9233	1st 2nd 3rd	TGTGATGCTGGAGACCTTGA CAGTTTTGCAACAGGAGCAC CACCCGTAAGGCTTTTCTCC	GCCACAGTCTGATGTGAGA TGCTAGCCAATCAGGACACA CCACAGCTCTACATTCCGTGT
chr19:117506242-117629361	123120	30	181.128	1st 2nd 3rd	CAAGCGAACAAACCCTGAAT CCCTCCGAGATCTGCTTATG CCTTTGAAGCTTTGCCTTTG	CAAGCGAACAAACCCTGAAT CCCTCCGAGATCTGCTTATG ACATCACATTCCGGGATCAT
chr19:126827482-127023260	195779	29	58.5965	1st 2nd 3rd	TCACGTCAACTCTTCCCTGA CTGCGTGGGGTTTCACTAGTT TGACACAGGCAGAGGCATAG	TGGCAACATATTTCCAAGCA CCTTGGTTCACTTGCTCTCC TTATGACCGAGCCCTCTTTG
chr19:65699679-65934932	235254	23	150.713	1st	TAGGAGGCAGAGCTGGAAAA	CAGCTCGACTAGTGGCTTCC

Deletion CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chrX:65699679-65934932	235254	23	150.713	2nd	GGCTGGTTCCTAAGATGCAA	CCATGGATTGCCAAGAAAAT

**Table A5.** Primer sequences used for validating 87 duplication CNVs by qPCR with SYBR Green I

Duplication CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr1:54756695-54842677	85983	51	51.25	1st 2nd 3rd	CCAGAATGTCGGAAATCACC GGCTTTCTTCTGCTGACACC AAGGTCTGTCCAGGTTGGAA	TGGAGAGTGTGAGGAGATGC AGAGGACCTTGAGGCAGGAG TCCAGGACCACAAAGGTCAT
chr1:103166045-103200900	34856	27	54.4936	1st 2nd 3rd	CACCAGAAAACCTGGTGAGAA ACGTGTACAGGGTAGCAGCA GGACCTGGATCACCTAAAGA	CTAGAGGACAGCAGGGGATG AGAACGTGGGTGAGCAGGTA GCATTCCACCAATATTCTCA
chr1:243187849-243498562	310714	140	260.502	1st 2nd 3rd	GTCGCACAAACAGGACAGG TGTGGGGCTGAGGTATTTTC TGCTTAGCAAAACCGTACCC	GGCCTAGCATGTTGCTTCA GGGTGACAAAAGCGAAAGTC GTGTTTCCGAAAGGCTGAAA
chr1:244601075-244774481	173407	108	189.364	1st 2nd 3rd	GCGGTGCTTTTCAAACAAAT TCTTCTTCTCGCACACAGA ATAATGAACCGCTCCACCA	AAGTATGGTCCGTGGACAGG TCACGTGAAGGTGTTTGGGA TGCACCCTCAAGACTTCAAA
chr2:47985440-48507207*	521768	99	34.4042	1st 2nd 3rd	CCCCACCTACCACTAGAAA GCTTGCTACCAGAACCAGCA CATCATCCCACTGCTGTACG	ACCCATCCCTTTGGCTTCT GGAAAATGGTCCAGAATCCA CAGGGAAGACACGCTAGACC
chr2:56247422-56277102	29681	20	52.1844	1st 2nd 3rd	GCTTTGTATTGGGCTGGA AGGACCTGTCAAAGTGTCG CCCATCACCGAATGTCTAC	ACCCAAAATTGCAGTCAGC TGAGGTTGCTGTGTCCAG CTAGCTAGCGCTGCATTTT
chr2:99205971-99283049	77079	29	68.9621	1st 2nd 3rd	TTTTATGCCCCCAAATCCA TAACAGGAGGCAGCTGGAGT ATCCAGCCCAGAAATGTGAC	GCCTCCTGTGTAGCAGGAAC GACCCCTGACCACTACCTGA AGAGCAGGAGACCATGCCTA
chr2:130539728-130864578	324851	108	147.542	1st 2nd 3rd	GTACCTGGTCAGCACCCCTA AAACAGGGCAAATGAACTGC ACTGCAGGTTTTGGAAATGG	CCACGTCAAGCTTCCAGTCT CGACCCTGGGTAGGTAGGTT GGTGAGCAAGGGTTCTCAAA
chr2:220068927-220112970	44044	25	42.6615	1st 2nd 3rd	CGTGCTCAATGCTGATGTCT ACCCAGAGCCTGACATGGTA GCTGCCACATAGTCGGAGTT	AGCTCTTGGGATCACCTCCT TGGCAGAACTACCCACAGAG CTTCTTCCCATGCATTTCC
chr2:220125968-220188031	62064	38	59.2593	1st 2nd 3rd	AGCCTCGCAAGGTTAAGATG CTTGGTCTCCCTCTGCTCTG TCCTGCTGGGTAGATTCCAG	ACGTTCCGGTGTGAAGTCTC GAAACTGGGAGGGTACACGA GGTCAAGGGCAGGACTTGTA
chr2:242322777-242393086	70310	40	31.1488	1st 2nd 3rd	ACACGGGCTATGACCTGAAG TGGTAACCTGCACCTCAATG TTCAACATCATGTGCAACCA	ACCCATCTCAAGGCACACAT AGGACGTCCTCTTCTCTGA CCCCAAGCCTTACACTCCTT
chr3:1285741-1575422	289682	236	382.449	1st 2nd 3rd	GTGTTCAAGGTCCACCCACT AGCACCCATTTTCTTCCAT TAAATGGCTCTCACCGGAAG	ATTTGGCTCCACCAACAAG CTTGTGTGAGAGGTGGAGA GGAGGATGTCCCTGTTCTGA
chr3:40985257-41512416	527160	187	305.322	1st 2nd 3rd	ACGCTATCATGCGTTCTCCT ATGGCTGAGAGTGTGTGCTG GGCCTGGAACAGAGAGAAGA	CTCACGATGATGGAAAGGT GGAGAGCCTCAAGATGCAG TTCCATCCTTGGGAACTGAG
chr3:143289345-143568000	278656	75	85.1647	1st 2nd 3rd	TGCTCAATCCATGTCAGCAT CTTTTCCCCACGAAACAAA AGCAGCACACACAAGTACACC	TCAGCTCTGTGATGCCAAGT GATGTGTTGGAGTGGGATGA GGGTGCAGTTCTACCTCAAGA
chr3:196055397-196587456	532060	230	393.975	1st	GGGACAAATGTGCGTAGTGTG	AAGCCAAAGGCTCTGACTGA

Duplication CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr3:196055397-196587456	532060	230	393.975	2nd 3rd 4th	TAGGCCTCGAAGACGTCACT CGACCGAGAGGAAGAAGATG TTGTTGCTGTCAGAATGTTGG	GGACCAGGACTTCTTCACCA CATCTTCCACGATGTTGCTG AGATGCCAAGAAGCAATTCG
chr3:197873977-198161806	287830	115	289.098	1st 2nd 3rd	TTTCAAGCCTTTTGCCTGT TCCAGATTCGTCAGCTCCT CTGGCTGAATGACCAGGTTAG	AAAACACTTGCAGCCCTCAC TGGCCAAGGAATCTTGAAAC CCTGGGGCAAGTAACCAAGTA
chr4:144333863-144496222	162360	37	118.866	1st 2nd 3rd	GATCTTTCGCTTGCCTTTTG TTTCCAACCTCTGGGTGAGG CTCGCGTTCTGTTCAAGTTTC	TTCTGAAGGACCGGTACAG ACTTGCAGGAGCTGGTTAAGA GGCGAGAACCTAGCTCTCCT
chr5:118487627-118649571	161945	57	104.324	1st 2nd 3rd	TTGCTGACAACATCACACCA CCCAGTCAGGGCTTACAGTC TCAGCATTTACCTGCCATGA	CTGCACTGGGATCCTGAAAT TTTAGAGGATGGGCCACAAT AAGCTGCCTCTGTTGTCGTT
chr5:180379341-180448334	68994	38	34.7147	1st 2nd 3rd	AGGCCACAACAGACTTCAGG CCCTTCAGAAAAGCTGAGGA CCTGGCTACAGCCCTATGAG	GGAGACTGAGAGGTCCACCA CGCTGGAATAAACCCCTCGTA CTTCAGACGGAGCAGGAGAG
chr6:84547600-84691015	143416	39	118.216	1st 2nd 3rd	GGAGGTGGGTGATAACTGGAT GAGCAGCAGGATCAGATGGT CGTTTTGTCTTGAGGGATTTG	AGGCCTGTCACTCACCTGAC GCCAATGCATCTGGAATTTA CTTCCTGAACCTCATGGCTTA
chr6:84847263-85138189	290927	89	210.706	1st 2nd 3rd	CAAAGCTTGTCACCAGACCA CATTTCTCTCTGCCCTTTG TCAACGGTGCACCAACATAC	TCACAAAGTTGGGGATGTCA CAATGAAGAACTGGCTGCAA CGAAATCAACAGCCTTTTCC
chr6:96654375-96715587	61213	27	52.6145	1st 2nd 3rd	GATTGCACTGGCAAATGATG AGGAGGAATGCAGGCTACAA GCTTCTGGTACTCGGTGAGG	TGAGATTAGCCGCCTGTTGT AAGCTGGGATTCCTAAAGC TAGTTGCACTGGCTCACAGC
chr6:144506912-144807770	300859	98	212.135	1st 2nd 3rd	ATGACCAGCAGTCCCAGAC GATGTCAAGCTGAACCATCG CGCTGAGCTTCACTTCTGTG	CAGCAGCTGGTTTTATCCT AATTCGTTCTGCCCATTTGTC GCATGTACCCTTGTGGAACC
chr7:1353711-1531935	178225	89	195.713	1st 2nd 3rd	GACCATGATCACGTCAATTGC GCAGCTGCCTCTGTATCTCC TGGGAGATTCTGGTGAAGG	GGGGTGAGAATCGAATGAGA TCCCCAGTCAGGGTGAGTAG CTCTTCACGCATCAGTCCAG
chr7:21649161-21908645	259485	207	509.627	1st 2nd 3rd	CCATAGTGGCCTACGAGGAA TTTTGCTTTGCAGGTTGATG AGCATGTTGCTGCACTGTTTC	CTTCAGGGCTGTTTCGTAGC GCTGCAAAAGATTGGTTTCG CGCTCAGCCCTGAGTTAGTT
chr7:55095278-55433865	338588	180	440.403	1st 2nd 3rd	TCGTGTGCATTAGGGTTCAA GGGCCATTCTAATAGCCTCA TCCTAGAGGACGCTCTCTGC	ACATAACCAGCCACCTCCTG ATGAGGTACTCGTCGGCATC AGGTGGAGCTTCAGCCTCTT
chr7:100207480-100244080	36601	31	51.5787	1st 2nd 3rd	TCTCGTGAGGGAAGGAAAGA CACTGGACACCGACTGATGT TCTTTCTGCCAACAGTCCAG	GCTGGGCATCATTAGTTCGT TGAGCAGAGCAGATCCAG CCAGAGCTTCTTCCCTCTT
chr8:175887-277897	102011	81	67.1141	1st 2nd 3rd	GCCTGTGGAGATTGAGAAGC GGCAGAGATAGCCCTGTTTG TCAACACAAGGAAATTCACACC	CCTCCACACAAGCCTCCTAA GCTTTCTCTGGGATGTGTCC TTTCCACACACACGACATT
chr8:27522232-27652542	130311	60	108.328	1st 2nd 3rd	CTGCTCAAGGGTAGGGAAGA CGGAAGACAGACGAGGAGAC TGCCAACAATTCTGCCAATA	GCGAGCAGAGCGCTATAAAT GATGTTCTTGACCGCCTCTC ACTTCTGATGGCCTTTCCT

Duplication CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr8:52376255-52818072	441818	150	388.467	1st 2nd 3rd	TTTCTAGCATCTTACCCACA AACCCATGTCCATGCTCAGT GGGGGATGTCAAGCAAGTTA	GCCTTGATACATGCCAAATGA ATTACTTTTGAGCCGACAGGA GCCTCGTTCTCCAAGTAA
chr9:3739196-3835784 chr9:3931973-3995592 chr9:4011386-4196769 chr9:4201589-4366397	96589 63620 185384 164809	59 47 168 142	136.453 83.717 396.494 253.682	1st 2nd 3rd	GCCCATGGATGTAGGACTA TGCGGATGATGGTATTGAAA GAAAAGGGCTTCCAACTCC	ATGCTTGCCTTCAGGATTT CCTTTTGGCACGGAAAGTA TAAGCCCCAGGAGACAAATG
chr9:4400854-4489426 chr9:4495544-4537288	88573 41745	58 28	198.9 60.3107	1st 2nd 3rd	CTGCTGGTTCCTGCTCTACC TCATGCCCTATTCCAATCC CTTGGTTCGAGAACACAGCA	GGTGC GGATTAACAAAGAA AAGGTGGTAGAGGCAGCAGA AAGGTACCTGTAATCATGCTGGA
chr9:19316053-19376565	60513	37	82.1049	1st 2nd 3rd	GCTCATAAATTGGCGAAGAGA AATCCTCCCCCTGTTTCTGT GGGTAAAGACGCAATCCA	TCATATGCCTGCTGAAGTGC CGAGGTTCAGAAAATGATTG TCTGAAGAAGCAGCGTACCA
chr9:122779165-123328769*	549605	610	443.871	1st 2nd 3rd	AGACACAGTTTGGCCTGGAG CAAACTGTCAAGCCACCT AGTCCCATAGCCGAAGCTCT	AGGAGGCTGCCAAAGTTTTA AAGCAGATTGGAGGCACAG CAATCACCTTCCCTGCCTTA
chr9:131514006-131766096	252091	154	360.921	1st 2nd 3rd	CCACCACAATCTGGAAGGAA TGTCCTGTGACTCGAGCAGT GAGAGCTTTGCACGTCCCTA	TCTGCAGATCCTCTGGGAAG CTTCATCCGCTCCTTCTCAG AGCCAGATGCGAGTTACAC
chr9:139403162-139519804	116643	52	144.059	1st 2nd 3rd	CCATCAGGCTCAGACTCACC GCTCCTCACCAAGTCTGAGC CGGCCTCACCTGTATATTGC	TTCTGCAGATCCTGTTCTCT GGTGCCAGACGAATCTGACT CAGAATCATGCGGAAGCAG
chr9:139973769-140061878	88110	29	31.916	1st 2nd 3rd	TGTGAAGTCAGCGTCTGGAG AGATCCTGACGGGAGAGGAC GTAACGTGGACCTGGAAGC	AGAGCAGGCTGATGATGGAC AGCCCCACAGAAGGATACAG GTTGGTGGGGCTTAAACAGA
chr9:140089674-140151812	62139	33	59.4435	1st 2nd 3rd	CTTCTGTGCTCCTTTCTGG CAGTGGGCGCATCAGTTAC AAGCAGCGCTTCTACTCCTG	GACATGGCCTGGATCAGAGT GTTGCAGGCTCAGGCTCTAC GCCCTGACCTTCCACCTAT
chr10:27654698-27832689	177992	116	165.246	1st 2nd 3rd	AGGCTGTCGATCTGGAGTGT CCTTGGGTAGGTACAGGAAGC ACACAAACCCAGCAACAACA	ACATGTCACATGCCCTTC TGTCACACCGACTGCCTAGA CCACAAAGTGCTGTGCTAA
chr10:28599898-28728684*	128787	75	38.7701	1st 2nd 3rd	CAGTTCAGCAACTGCCAAGA GGAAGACAGCTGTGGAACC ATGGGAATGCCAGTGATTTT	GGAGACTTTTCGCCACATA TAGCTTCTCCAAGGCCAAAA TCCCTCTTTCAGGCTACAT
chr10:53735827-53798419	62593	23	60.2727	1st 2nd 3rd	CATCAGACTGTGCTCAGGA GGAAAAGAGGTCTGGGAAG GGGGTTTTGGGAGAAATAA	CCACAGTAACAACGCTGGAA CCTCTGCTCTTGTTCCTCA CCTGTCATTCTCAGCCTTC
chr11:5359130-5388870	29741	42	99.1799	1st 2nd 3rd	AAGGTGGTGCCACTTCTGTC ATCTGCCACCCTCTGAGGTA CTGGATATCCGTCCCCTTCT	AGAGTTGTGTGGGATGAGC AAAGCCTTCAGGAGGAGGAC CAACATGGCAAGGAAGAGGT
chr11:18357549-18396314	38766	23	43.0579	1st 2nd 3rd	TCACCCAGGAGAACCTGAAC TGTTCTGCTTCCACATCTGC TCTGTAATGATTGCGCCAAG	GCTCCTCTTTGGCATGAG AAAGTAGAGGGCACCCATT AAAGCAAGGCCAGCTACCT
chr11:108254505-108319661	65157	25	41.9953	1st 2nd	CCCAACCGTAGCTGAGAGAG ATATCCATTGGGGTCTGCAA	TAGCGGCTCTGGAATAGGAA ACTCCAATCCAGAAAGGCTTC

Duplication CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr11:108254505-108319661	65157	25	41.9953	3rd	GCCAGGTGGGACACTTTAGA	TGCCTACGCGATCATAACCA
chr12:95960-333509	237550	162	253.467	1st	CCTCTCTGTCCTCGTCCTTG	TTGCGGATTTTCTCGAAGTT
				2nd	GACTCACCTCAAGCTCCTG	CGTTCTCCTGGGACTGGATA
				3rd	CCTTTTCTTCCCATCGTTCA	TCAGATCCGCAACAAGTCAC
chr12:1071807-1311398	239592	78	125.37	1st	TGAAACAGGAGCTGTCCAGA	CCCTCTGCTCCTTAGCAGTC
				2nd	TCATTGTTTGGGACATGGTTT	CCAAGTGCTGTATTCGGTCA
				3rd	TCAATGTGTTCTCGCTCTC	GGCCATCAGTAACTCCTCCA
chr12:18459223-18485035	25813	11	35.4792	1st	CCTTAGTCACTGGGGTAGA	TGTTGAACTGCCACTTTACGA
				2nd	TTCTACTCCACGCTCCTTG	TCCTCTAAAGCTGCCCCATA
				3rd	GGTGATCAGACGGTGACTTG	CTCAGTCAGAGCCACACTGG
chr12:51330003-51442331	112329	109	120.634	1st	GGATGACGTTGGTGTCAAGT	AGGGTCTGCGTTTCCCTTAT
				2nd	GGCTCTAGGAGGCTCTGGTT	CTTTGGAGGAGGTGGTTTTG
				3rd	CAGTTGCAAAAGGTCCCAAT	AGTTCCTCAAAGCCAGGAT
chr13:24169434-24228338	58905	22	52.948	1st	ACTGAACCATCCCCACTTTG	GCTGAAATGATCCCCACACT
				2nd	CATGGTCTGCTGGGTGTATG	GAGGCCATAGGAGAGGATCA
				3rd	CCAAGGCTCTACTGCTGGAC	AAGCTGTGCGTTCTTCAAT
chr13:30626847-30772444	145598	47	53.1348	1st	TCTTGGGTCCCCACACTAAG	GCTCGCAGAGCTGCTACAT
				2nd	TGCAGCTGCTTAAAAATGAGC	CCAAACCAAAAGCTGGA AAA
				3rd	TGTGCCAGTGATTGTCACCT	AGGAATCCCAATCCACAG
chr14:22925866-22953123	27258	32	43.1433	1st	GGGTGGTGAAATCATTGAGG	TCACCCTTCTGTCTTGCTT
				2nd	CTCTCCCTGAGACACGAAGG	CCCTCCCTTTCCCAACTCTA
				3rd	CTCGGTTTCAGCAATGACCT	ACCTCCGCAAGTCAGAGAA
chr14:23985223-24057519	72297	34	44.7893	1st	CAGGAGACCATGCCTGAGAT	TCCAGGCCATGTTCTTTGAT
				2nd	AGGGCATTCTTTGAATGTGG	ACTCCTGGCTGTTGAGAGGA
				3rd	TTTTGTCTTCTGGGATTGC	CTGGGGAAGAACAGGTGTGT
chr14:31579294-31628653	49360	19	64.3384	1st	GCAACTACCCGATTGAGGAA	ATCATGCATTTGGGAGAAGG
				2nd	GATCCACGCTTTGGCTTCTC	CCCAGAAGACGAAACTCCAC
				3rd	TTCCACCTGGAAGGGTACA	GAAGAGAACCAGGGAGGAG
chr14:68992305-69046721	54417	13	33.1456	1st	GCCGGGACATTTCTTTATGT	GATGAGAGGGATGAGGCAAC
				2nd	ACAAGGGGATGCTGTCTGAG	CAGGCATTTGGATGTGTGAC
				3rd	ACAGTGCTTGCTTGGGACTT	GCAGCTGATGACACTTTGGA
chr14:103716079-103873344	157266	69	71.4027	1st	GGAGTGGGACTTACCTGCAC	ATGGATCTGCGTTGAGAACC
				2nd	CATCTCCTCGTGTTCAGAA	TGTAATGGTGCCCGTCTGTA
				3rd	AGGTGTGTCTTGGGAGGATG	CAGAGCAGATAGGCCCTTGG
chr15:88664833-88752520	87688	40	60.2516	1st	CACAGCCCTGTCACGAATC	CCCTCCTTGAATCACTACC
				2nd	CGGAAGCCACACCTCATAAT	GTGTGGGCTCCTCTGGTGAGT
				3rd	CGTCAGAACGTGGCTTATGA	CCACTGTGTTCCAGCCTCAAA
chr15:98367631-98710098	342468	217	312.989	1st	AAGTCGCCCTTACCAAGT	ATGAGGCCCTCTGTTCTTT
				2nd	CTGACATGATGTGGGACCTG	AGCACCAGGAACTGGAATG
				3rd	CACGTCGGCTTCTTCTTTTC	GCCTGTCGAGACTGACCACT
chr16:20465942-20630271	164330	48	71.4004	1st	TTGGGGCTGTTTGTTACCTC	AACTGCAGCAGCATGTCAAG
				2nd	TTTGCTGATAGGGGAAGGTG	CTGTCGGAAGTCTGAGTGGAT
				3rd	ATCTTGGGGCTCCAAATTCT	TGAGGACTGTTTCAACATGC
chr16:22980334-23025207	44874	19	52.6576	1st	CTGAAAATCGGGATGGAGAA	GTGAGAAGTGCCAGCGTCA
				2nd	TGAGCAACCAAGAAATGCTG	CCTGATGCAGAAAGTTCG

Duplication CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr16:22980334-23025207	44874	19	52.6576	3rd	GGTGGCATTATGGAACAGG	ACGACCTGGAAACAGTCCAT
chr16:79757966-79888598	130633	127	256.56	1st	CTGACTTCAGAGGCCAGCTT	CTGACTTCAGAGGCCAGCTT
				2nd	ATACCAGCAACTGCCCTCTG	CAGAATCCCCATTCTGCAC
				3rd	AGCCTGTGCAGACCAAGTTT	GTTCAGGTTGGCCAGGTAGA
chr17:3708409-3747829	39421	35	69.2656	1st	TGTTGATCCCTCCAGAAAGG	AGACAAGAATCCCGAGACGA
				2nd	CGAGAAGGGCTTGTACACAG	GACTCATGTGGGCACAGGTT
				3rd	ACGTGAGCACACAACACACA	CCGTCCTCGAAGTTTCTGAC
chr17:77925586-77986847	61262	26	49.3379	1st	CATCGTGCTGCTCTTCTGG	TAGGTGAGGCAGGTGGTCA
				2nd	TTATGCGGGAGAAGAAGTG	ACGGCTATTGTGTCTCTGTC
				3rd	CTCTCCAGCCCCTCTGAAAT	TGGCACCAGGTAACCTCCAT
chr18:723474-929338	205865	79	162.859	1st	TCAAATTGTGCTCTGGTTGTG	TTCACCAGGTGCTTATTCC
				2nd	TGAACCTGGCACCCTGAAA	TGTCACCATGTCCGTCATCT
				3rd	GGGATCTTCACGACAGCTA	AGCTGGTAACCCATCGCTAC
chr18:9066995-9126155	59161	23	52.5421	1st	TATGAGAAGCCACCCTCTGC	CCTGACGGCCTCTAAAATCA
				2nd	GGGAAGACATGTAAGGAATTTG	GGATTTACCCCAATTTCCAAG
				3rd	CAGGGCCAAGGTATGCTTTA	CCAAAGCTCTACCAATTCC
chr18:30679496-30915415	235920	78	129.858	1st	TCCCTGAAGGTTCAAGAAGAA	CCTCCATGATGAGCAGACAA
				2nd	GCTTGAAGTGGCATGACTGA	ACTGCTTGAATGGCCTAGC
				3rd	AGGGATTGACAGCAGAGGTG	CGGTGGCCATAAACTCCTTA
chr18:31152359-32009041	856683	279	470.034	1st	TGGAAGGCTCTAGACAGCTCA	TGCCTTCAATTCAAGCACTC
				2nd	GCCAATAGGCCATTTCCAGAG	TGATTCGTTTCCACATTTG
				3rd	TGTGCAGTGAACATCCTTGC	TCTCACTGGCTGCTGTAGGA
chr18:32045011-32098386	53376	24	55.0338	1st	TAGGCTGGTCATGCTGTGG	GTGCAGGCGAGTGCTATGA
				2nd	GAGGTGCCTCTTGAGGAAAA	TTACCCCTCTCCCCCTCTGT
				3rd	GCCATCCTCCCCCTACTTAC	CTTCAAAGCCCTTTTCCA
chr18:75342653-75507409	164757	136	198.576	1st	TGCCTGATCACAATTCG	GGACTTAACCCCTGGCTCAC
				2nd	TAGTCGGGGTTGTTGAAAG	ATAAGCCCCACGTGTTTCTG
				3rd	GTTGGATGCCTATGGCACTT	ATCCCAATGCTAACCCTGC
chr18:75515931-75559025	43095	20	64.4395	1st	TCCTGGGTTATGGAATCGTC	GCTGAGGGAGTTAGCAGGTG
				2nd	CGGTCCTGGAAATAAGGACA	GGCGTCAACAGGGAAGAATA
				3rd	AGCTGAACAGCTGGGAAGAG	CGACATCTGCTGACAGTGCT
chr19:61150873-61219157	68285	67	92.4029	1st	CAGAAATTCCTGGGTCTGT	CGAGAGAGGGACATTTACGC
				2nd	CTCCGCAAAAACCAACATCT	CCCGGTGGAGACTTACAAAA
				3rd	CGCTCACCTTTTCCAGCTAC	ACGTTGGCATTCTCGATTTT
chr19:61413858-61898207	484350	298	316.716	1st	TGGAAGAGCCTTCACTCAAAG	CCAATGTTCAATCAGGGAAGA
				2nd	AAAACCTTCAGCTCGGTTTC	TTGATGCTGAGTGAGTGATGC
				3rd	CAGGGCTCGATATGTTCTT	CAGGAAAACAGGAAAGCTG
chr20:828964-979043	150080	166	183.098	1st	GTTGGAGTCCCTGGAGACCT	CACCATGTCTGTGGCTCAAC
				2nd	AGCCACGTACTTTTGCACCT	CAGCAGAGGCTCTTCTGTT
				3rd	CTGACCATGCGGCTACTACA	CCTCCAGACGAGTCTCAACC
chr20:5352295-5718919	366625	149	272.438	1st	GTTTTCCGGGAATTTCTGCT	GCACGTAGTTCAGGGTGATG
				2nd	TCAGTGACCTGCCACTTCAG	TACAGGCGAATGCCACTATG
				3rd	CGAGGCTTTCAGTTCTTTGG	ATCACACACCCGATTGCTG
chr21:46515232-46536901	21670	20	44.2269	1st	CTTTGGAATTTGCGTGGATT	GGACTACCTGGTGACCCAGA
				2nd	AAGCAGAAGTGCTTGGGAAA	TGTTGGACCCTTTTCTGGAC

Duplication CNVs	Length (bp)	No. Probes	Max. Log BF	Primer pair	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
chr21:46515232-46536901	21670	20	44.2269	3rd	TTTGCCCTGGGATTAACCAG	GAAATCAGGCTGGGGAAAT
chr21:46540022-46625020	84999	34	121.771	1st	CCTTCCTCCAGATGTTCCA	GTCCTCGCTTTCAGAAAGAA
				2nd	TTCCTGACCGTTGAGTAGGG	GCTAATCACCTGGCTGTTCC
				3rd	CCAGAGTCTTGGGTCTGGAA	TCATGCTGCAACTTCCATA
chr22:28615613-28885216*	269604	80	52.359	1st	GGGTAGGGAGGAAGCAGTTC	ACTGTCCACGTTTGGTCTC
				2nd	CACCTCCGTTGTTTTCCAG	GGAAACAGGCCCTTAGGTA
				3rd	AACGTGCAAGTGGGATTTGT	ATGGGCGATCTCAGTAGTGC
chr22:43955895-44001348	45454	40	43.1688	1st	ACCACGAGAGAGGCTTTTGA	CCCGGTTCAATTGGTAACAG
				2nd	AGAGCTGGCTGTGAGAAGGT	GCTGAACAGAGCAGATGGAA
				3rd	GCGCACATAGAAAAGCATGT	CGTCCTCCTCACTCCTGTC
chrX:6016157-6169284	153128	37	87.8562	1st	GATGATGGTGCCTTGATCCT	TGGCTAAGTATGCAATCCA
				2nd	CAAGGAGACAAAGCCCAGAG	AGAAAACGGCTGTGCTTGTT
				3rd	GCGAGACAGATGGGAGAAAG	TTGACTTGCAGCAAAGGATG
chrX:18689860-19105471	415612	72	85.1306	1st	TCCCAATCGGTACAATCGTT	TCCCTGAGAGAAGATTCATGC
				2nd	TACTGAGCATGGTGCGACT	GCCCTCGACAATGAGATGAT
				3rd	TTGAGGCTTGTGTTTCAACC	CCTCCTCCAATGGTACTCCA
chrX:46443751-46604442	160692	24	33.1653	1st	GGTGCTTGAAGAGCCAGATG	ATGGAGGAGCTCGCTACTGA
				2nd	GGCTGCTTCTTCCAAGAG	GCGGCTACACGACTTTCATT
				3rd	CCAGAAGATGCTGTGGTTCA	TCGCTGCTCTTCTGTCTCTG
chrX:117664921-117848187	183267	41	35.3568	1st	GCCAAAGAGCTTGATCCAAA	AAGGGGCTCAAAAACAAAT
				2nd	AACTCAGCCACCTGTGACAA	TTGTTTGTGCCCCAAATGAC
				3rd	GACGACTCCGATGAGGATGT	GGGAATCAATGACCAGCACT
chrX:118311796-118492016	180221	36	71.2488	1st	CGGCCCATACCATATTTCTG	TGACGGACACGAAGCAGTTA
				2nd	TTGGGGATATCGTGCTTCTC	CCTGCACAAGGTTTTCTGT
				3rd	CACCCAGGCTCTTAACCTCG	CACAAAACACAGGGATGTGG
chrX:152871559-152945374#	73816	31	30.7141	1st	GACAGAGATGGGGCTCTTCA	TCTACCCGCCCTATCATCAC
				2nd	TCTGTGACGTGCTCCAGTTC	GTCATCGGTTTTACAGCAGGT
				3rd	GGTAGAAGAAGGGGCTGGAG	CCTGCCTCTCACTGGTGTCT



**Table A6.** Primer sequences of reference genes used for CNV validation

Reference Gene	For Primer (5' – 3')	Rev Primer (5' – 3')	Product size (bp)
<i>BNC1</i>	TCAGTGCTTTGTCCAACAGG	GCAGATGTCACACTGGAAGC	125
<i>CFTR</i>	GGAGATGCTCCTGTCTCCTG	GGGAGTCTTTGCACAATGG	138
<i>RPP38</i>	CTGCCATGATCACCTCACAC	GAACGCCAAGGCTAGAACAC	128

**Table A7.** Primers used in Sequenom analysis. The SNPs were divided into plexes 1 through 4 containing 37, 34, 30 and 17 SNPs, respectively.

**PLEX 1**

SNP ID	Forward PCR primer	Reverse PCR primer	UEP primer
rs10823418	ACGTTGGATGGTTGGAGACAAATGTGGAGC	ACGTTGGATGGCAGATACCCAAGTTCTTCC	agGCTCACCAACTCTTGT
rs1512414	ACGTTGGATGGGAGGGAGATAGGATCTCTA	ACGTTGGATGTACCCTACAGTGCCTGAAAC	AACCGTCACCATGAG
rs1799932	ACGTTGGATGGATGCTGTGGATGTCAATGG	ACGTTGGATGATGAGTCCCTCCCCACCCT	ttcCCTTGGGGCACTTCA
rs2112613	ACGTTGGATGCTGCTTACATGGGCCCTTGTG	ACGTTGGATGAGAAGCTATGAGCACATGCC	GCCCTTTCACATCCTA
rs292488	ACGTTGGATGCTCTGGTCAACACCCTTTTC	ACGTTGGATGAGAGAGAAGAGGTGGGCAAG	GGTCAGCAAAGACAGA
rs667808	ACGTTGGATGAGGGAGAGAAGCAGGAATCG	ACGTTGGATGACTGGTGAGTCACTCTTGT	gTTGGGGAAGTGGCCAC
rs7727544	ACGTTGGATGAGTGGCTCAGGGCACGATG	ACGTTGGATGAAGACACAGTGAATCTTCC	AGACCTGGCTGCATA
rs7920199	ACGTTGGATGCTCTGGAATGACTTCTCCG	ACGTTGGATGCTCTCGAAACATTCTCAC	ccTTCAGTTCCCGCAC
rs8098464	ACGTTGGATGCTACAAAAATCCATCTAAAG	ACGTTGGATGCATGGAAGAAGTACTGAC	AGTGTACTGACTACCT
rs10848666	ACGTTGGATGCATTGCTCACCAGGCAGT	ACGTTGGATGGAAGCAGGAGGACAGATGAG	ACAGATGAGAGATGGAGA
rs11645638	ACGTTGGATGCAGAAATTCAGTAACCAAGAG	ACGTTGGATGAAGAAAAATAAGCCTGGCCC	aTGCCCCATGCAATTTGAA
rs17151639	ACGTTGGATGACCTGAGAGATGGGAGAAAC	ACGTTGGATGCGTGCCTTATCAGAATGGAC	agCCTGTATACTGGCAGTTC
rs1887432	ACGTTGGATGGAAGGCAGAGCCATTCCTAT	ACGTTGGATGTTTTGACAAAGGGCTGCCTC	ccccGTTGGGCTTGTCT
rs2158641	ACGTTGGATGTCATACACAGTGGCTACACG	ACGTTGGATGCTTTCCTACCTGTTTTTCCC	TTCCTGGCACTTCTATAGC
rs2274170	ACGTTGGATGCAATCATTATGTGATAATGT	ACGTTGGATGAAAAGTACCTGGGAGTGTG	agGGGAGTGTGGCAAGC
rs2304644	ACGTTGGATGCAGAGAAGAAATGTGGCAAG	ACGTTGGATGCTCCATAAAGTAACAGCAAC	AAGTGAAAGGAGCCTAGTT
rs6884552	ACGTTGGATGCACTGCGCCAGCCTAATT	ACGTTGGATGCTGCAGAGCTGAGAAAATTC	tttgGGCATGTAAGGACAGA
rs7191411	ACGTTGGATGCCCCGCCCATAAAAGAAATG	ACGTTGGATGTCATGGAGAAGAACGCTCCT	ACGCTCCTAAATGTATGTAT
rs12139641	ACGTTGGATGGAATAAGTTGGAGACTTCCC	ACGTTGGATGTTCCACACCTGAGCTCTG	ggggCACCTGAGCTCTGCGTTGT
rs16957347	ACGTTGGATGCACAATAACCCCTGGGTC	ACGTTGGATGAGACATAAAACGTTTGCCC	aatcTGCCCAACAGTTAATAA
rs17346550	ACGTTGGATGTCACATGGTATGTAAAGTGC	ACGTTGGATGGGAGAGTCCATCCATGTATT	ttcTTGTGCATGTTCCCTACC
rs2089855	ACGTTGGATGTATATGCTCACGCACAGAGG	ACGTTGGATGAACCTCCTCCTTCTCATGTG	tggGAAAGGCCAGATCCACACGCT
rs330297	ACGTTGGATGACTTCTGAATCCTGGACCC	ACGTTGGATGCAAGGTACTACTTCTGATCC	agacCCCTGGGACCAAGGGATT
rs6475895	ACGTTGGATGCCTATAACATACTCATCCG	ACGTTGGATGGTGAGAAAATAACATACTC	ACTACTGTAACAACACTATGTC
rs6981465	ACGTTGGATGACTGATTGCAGACTCTTCCG	ACGTTGGATGGCGGAATCATGATGCCATAG	ccccGCAGCCATTTTCTTTAGC
rs7997539	ACGTTGGATGGGATAATGAAGATATTAGC	ACGTTGGATGACAGGTAGTCTACAAAGGTG	cttCAGGCCTAAAAATAAGAGAT
rs942876	ACGTTGGATGGTCTGAGGCTCTTCTCTTCA	ACGTTGGATGTCCTCAAAACTGGGAAGAG	GGGAAGAGTTTTATAAAGGAA
rs11629255	ACGTTGGATGTTCTGCTGTACCCAGTTACC	ACGTTGGATGATGTGGGGTGTATATAGG	ctatGAATACCTCTTCTCCAGAGA
rs11730575	ACGTTGGATGCCAGGTTGTCAACCCAAAAC	ACGTTGGATGTGGTGCTAGACTTGGAATG	ccccTGGAATGGGATTAGCATATT
rs2400940	ACGTTGGATGCAGAGTTATGTCTCTCTTTC	ACGTTGGATGCCTTTAGCTAACGTGATCAAG	gacgTTAGCTAACGTGATCAAGAAAAGC
rs389557	ACGTTGGATGAGCAACATTGGAAGCAACCC	ACGTTGGATGAGTATTCCTTGCAGGATGG	gttgGCAGGATGGTACATTTATTTA
rs4860701	ACGTTGGATGAAGAGTATAAGGAATCCTC	ACGTTGGATGAACAAGGATGATGATCACG	tcttcTGATTGTTTTGGAGCTGAGTGA
rs544276	ACGTTGGATGGTCCCAACTCTGTGTTTG	ACGTTGGATGTGATTTGGAGGCCTGTACT	gcaaTGAAATTATTGGATTGAGTCT
rs60455014	ACGTTGGATGGCTTTGAGCTTTGAAGTAATC	ACGTTGGATGGACAGGAAAAATAAAGGAGT	CAGGAAAAATAAAGGAGTTAAAGTT
rs6997421	ACGTTGGATGGTTTCTAAGATACATTGTC	ACGTTGGATGCCCTCACATTAAGAAAGCC	gaagCAAAAGAGGAAATGTATCACAGA
rs731326	ACGTTGGATGCAAAAGGTAAGTTGAGGCC	ACGTTGGATGAGAAGTGTATCCATACTGAC	tAAGTGTATCCATACTGACTACTTAAT
rs927596	ACGTTGGATGCCTTGCTTTTGAGCAACG	ACGTTGGATGAATGGCGGAAGGACATTAA	GACAGTGATAGAAATGACTATTACTTA

## PLEX 2

SNP ID	Forward PCR primer	Reverse PCR primer	UEP primer
rs11012878	ACGTTGGATGTGCAAACTAGTCTGCTGCTG	ACGTTGGATGGAGGGTAAATTTACCTCAAG	ggCTCAAGAGTCTTGGGC
rs11158362	ACGTTGGATGATGAAAGCACTCAGCAAGCC	ACGTTGGATGGAGTGAAAATTGTCAGGAGC	ACTTCCTGCCACCCC
rs12935619	ACGTTGGATGCGAAAAAATATGAACTCTG	ACGTTGGATGGTTATTCTGTTACAGGTTG	TGCAAAATGCTCAGAAC
rs4392152	ACGTTGGATGACCAGCTTTTGGACATGCC	ACGTTGGATGCATGCCCATGTCCTTCAATC	TGCTCAAACCTGCAA
rs4789409	ACGTTGGATGTGGATAAACCGAAGAGGGTC	ACGTTGGATGAGGCATTGCTGACATCCTTC	AAAGGTTGGAAGAGGAG
rs624350	ACGTTGGATGGGCTGAATTGTTTGTCTCC	ACGTTGGATGAAAGCCTCAGTCCCAAGAAC	tCAAGAACTGCTCCCCCT
rs798379	ACGTTGGATGTACAGTCTCCCAACAG	ACGTTGGATGGAAGAATGGAATCCAGGGTC	CAGGGTCATGTGTAGG
rs9805437	ACGTTGGATGTGTTGCTGTGAGGACTGACC	ACGTTGGATGTTAAAATGTACCCAGCACCC	CCAGCACCTGTATGA
rs10065787	ACGTTGGATGTCAGCTGCACCTGCATCTT	ACGTTGGATGACCTTTCTGGGTGACTCAAG	cggTCCACCGTGTGTTGAAT
rs2169123	ACGTTGGATGGGAATGGAGAACTCCAGAC	ACGTTGGATGCTTCAAGAATGAAACCACTG	ATGAAACCACTGACCCTTA
rs2275199	ACGTTGGATGACATCACCTGTTTCCCTC	ACGTTGGATGTTTCCATGGTGTGACCTGGC	gAGGATCAGATGTCGCAG
rs2965228	ACGTTGGATGGTCCATCTCGGAGAATATGC	ACGTTGGATGACGGAAGAAAACTGCACCC	tttaGCACCCAGCAAAAAC
rs3111779	ACGTTGGATGGCAACAAGAAGGAATGAGGG	ACGTTGGATGTCAGATGCCCTGGATCCTTC	ccctTTTCAGACACTCTGACG
rs4236978	ACGTTGGATGAATGGTAGATGTACCATGT	ACGTTGGATGTACCTACATTTCTTGTTG	GGTTGAATACACTGAAACTAA
rs4968046	ACGTTGGATGTCAGGTGTTGCATTTTCCG	ACGTTGGATGCAAGATGAGGGAACATGAGC	gggtGGCATCGTGGAGCTT
rs6061772	ACGTTGGATGGACAGGAAGAATCCCATGGT	ACGTTGGATGGCATCCAACGCTGTAATCTC	ctCTGTAATCTCACATCCTCA
rs1329428	ACGTTGGATGGAGTGCCTAACTTTTACAAC	ACGTTGGATGATTTTGTGCCCTCTACTCCC	ggaaTCTACTCCCAGAACTAAGAG
rs13419910	ACGTTGGATGGATTTATTCTACAACATCCCC	ACGTTGGATGCCTGAGTTGTTATCAAAGAGC	gAAGAGCAGAATTTTCCAGATCAT
rs16931374	ACGTTGGATGTGCTGGATTTAACAGCTGAG	ACGTTGGATGGGCAATGAATCGGAAGTATC	ATCGGAAGTATCTGTCTCTGTG
rs2828064	ACGTTGGATGACATAAATACCAAGTTGAG	ACGTTGGATGCAGCAGCACCTCAAACAAAC	aAGCACCTCAAACAAACTTTAAG
rs6570786	ACGTTGGATGGTAGAGTTGTTACAGATTG	ACGTTGGATGGTTGTGCCCTTCACTTTTG	ccctCCCCCTTCACTTTTGGATAAC
rs7521700	ACGTTGGATGACTCTAGGGCAGGAAAAGAC	ACGTTGGATGGGTGCTCAGTAAATGCTAGG	agAGGAGTTATTTCTGACTAGG
rs797517	ACGTTGGATGGAGAGTTCATATCTTGAATC	ACGTTGGATGTGCAGACATAGTTTCTCTCC	cccgAGTTTCTCTCCTTCCATAG
rs9984896	ACGTTGGATGTAGTTATTAACAGAGAGCG	ACGTTGGATGTCTGGTTTAGGCTACCCAAC	catcTGTTGGCCTGTTTCCCTT
rs10153396	ACGTTGGATGAGTTGACTAGTTTGTACTC	ACGTTGGATGCAGATATCTTAGACCGCATC	ctccGATATCTTAGACCGCATCAATTTG
rs11161353	ACGTTGGATGGAAGGGCATGCCAAGGAATC	ACGTTGGATGCGCAACCTCTCTGTTCAAAC	ccccAACCTGCAATGGCCATGTATTC
rs13357903	ACGTTGGATGCTGAATCCACACCACTGTAG	ACGTTGGATGGAGAACAGGCTGTAAGAAATC	agaATTAGGAGGTAGATACTTCAGTCC
rs158896	ACGTTGGATGTCAGAGAGTTAGAGGAGGTG	ACGTTGGATGTGCAAAACAGGAGTTGATG	ggacAACCAGGAGTTGATGGTAGAA
rs16931326	ACGTTGGATGGGCATTAAAAAATCTGCAC	ACGTTGGATGTTCTTCAATGACCCTCCAC	CAGTATTTATCATGTGTTAAAGTCAT
rs16932506	ACGTTGGATGGAATGTCCACCACAGTAGTC	ACGTTGGATGAGGAGAAGGACATTGTCATC	cTTGTCATCTTCATTAGATTGTTTAT
rs17835866	ACGTTGGATGAGACCCGGAATTGTCCCAG	ACGTTGGATGACTCAGCTGATCAGCTTTGG	ggcgGATCAGCTTTGGTTTGATGTT
rs2918006	ACGTTGGATGAAATCCAAAAGTGGTGTGC	ACGTTGGATGTGGCAAGGGTTCTGTGAAAC	tcacCCAGTCACCAATATGCTAGTA
rs443685	ACGTTGGATGACCTTGGACATGACTGTGCG	ACGTTGGATGGTGAAGACAGAGAAAATCTAC	gggcAGACAGAGAAAATCTACCATACT
rs7569783	ACGTTGGATGGCTGACAAGTTCTGTATTGG	ACGTTGGATGCCAGATGTCACATCATGTC	ccTCATGTCAGTGATTTTTAAATTGTC

### PLEX 3

SNP ID	Forward PCR primer	Reverse PCR primer	UEP primer
rs1864262	ACGTTGGATGTAAGGGTGGAACTGAGAG	ACGTTGGATGGTAGTTCCTTTTGGTTGTC	ATCTCCAACCCCCTA
rs195656	ACGTTGGATGGAGTAGGAGCTGTGGTGCAT	ACGTTGGATGAAGCCACCATTGCCTGTTAC	CCTGTTACTGTGGTCA
rs2073167	ACGTTGGATGGGCAGGACGAGAAACAAAAG	ACGTTGGATGCCAGGAGATGAGTCTCCTTC	TCTCCTTCTCTCACCA
rs4858100	ACGTTGGATGCTAGACACCTTACCTGCCTC	ACGTTGGATGCTTTGCACATGCTGTGATCC	TGATCCCATCTAGGACA
rs6089491	ACGTTGGATGCCCTTCCAGAGAAGCATATC	ACGTTGGATGATGCACCATGCACATTTCCC	CCTTGGGGCAGTGTT
rs6536530	ACGTTGGATGGCCAGCAGAAAGAATCTGAC	ACGTTGGATGTCAACATATAGCCTAAGCAC	ATCCTCAGCAAGCCT
rs912284	ACGTTGGATGACCCAAATCACCACTACCG	ACGTTGGATGCTGAATGCAAAGCCAAGAGC	cAGCAGCAATACCGACCT
rs11578307	ACGTTGGATGCATGCAAGTGGCCAGTGTTG	ACGTTGGATGTTCTCCACCTGAATCCCAGT	GGCAGGGGAGCTATGCCTCA
rs11955347	ACGTTGGATGTGCCTTGACCACTCTATGAC	ACGTTGGATGATGTTTAAGGCCCCAGCTTG	GCCTACAGGAAAAACCTT
rs12597756	ACGTTGGATGGGGAGGTATAAGACATCAATC	ACGTTGGATGCATCACCCCTGCCTTCCG	CAAACCAATGTTTTCTTCAA
rs1351805	ACGTTGGATGAGAGAGGGTGAACTTGAGC	ACGTTGGATGTTCTCCTCACTGACTCATTCC	CAAAGCAACTGTTTTCAA
rs234434	ACGTTGGATGCTCAGCTGAGCTAGACCATA	ACGTTGGATGGAGGTAAAGCATTGCTGGG	AAGGTTTCGATAAGTGAT
rs8192120	ACGTTGGATGCCCTCCATCTTTGTAGTTAC	ACGTTGGATGCCCTCTCTCCTAAGCATA	CTCTCCTAAGCATATCCTAC
rs822431	ACGTTGGATGGACCCTTGATGCACTCTTG	ACGTTGGATGTGGAGGGCTTCTCGCCAC	GTGCAGTATCAGATGCAGTTC
rs1020430	ACGTTGGATGGATGAGTTCAAAGGCCATGC	ACGTTGGATGTCTTACTCTCTTGTCTCTC	ACTCTCTTGTCTCTCATGAAGTG
rs10770675	ACGTTGGATGAGAAACGGAACCTGGCAAGC	ACGTTGGATGTTTTCAGGTGCGGTTCTTAC	ggCTTACACCATACCCATATTTTC
rs1244229	ACGTTGGATGGCTTCCGGAATAAAGTGTG	ACGTTGGATGTTTCTTCTCTCTCTTTTG	aCTTTTTCTTTCTTTCTCTCTTC
rs149476	ACGTTGGATGGGCTGCTGGGGATTGTAGTT	ACGTTGGATGTTGTAGTCCACCTGCACAAC	agGCCAGAGAGCGGGGCCTCAGCA
rs330295	ACGTTGGATGGGGTCTCTGGGGCCACCCT	ACGTTGGATGAAACCTGAGGGTAGCCATAC	CATACGAACCTTTGAATATGTA
rs35407548	ACGTTGGATGAGGAACCAAGGAATTTGGGC	ACGTTGGATGTCTCCTTGAGGCACTAGGG	AATCTATTTCTTGTGTTTTCT
rs8192166	ACGTTGGATGCATGTGGAAGTATGTCAACG	ACGTTGGATGAGTGTGATCTGTCTATGAG	TGTTGATCTGTCATGAGATTTTA
rs10501538	ACGTTGGATGCTTCAGCTTTTAAGTTTCTGC	ACGTTGGATGGTTAAGAGACACCTAGTAAG	cTAAGAGACACCTAGTAAGAGCTCAAC
rs12015040	ACGTTGGATGGATTCTATCCCACTGACGCC	ACGTTGGATGGAGTTCATTGGGTTTCATAG	gaATTGGGTTTCATAGTATGGCAGTG
rs1354876	ACGTTGGATGTGATGTAGACTTTTGCACAC	ACGTTGGATGTGAGATATTTGCTCACCCAG	ggCTCTAGTCTTTAGTTCTCTTCT
rs1529017	ACGTTGGATGGGCATTAGGCAAAAGTCCATC	ACGTTGGATGATGTCCATGCAGGAAAAAGG	cTTGTCTTTTTTGGCCAAATGCCT
rs292489	ACGTTGGATGTTAGTGGTAGCAGCCGTTAG	ACGTTGGATGTACAGAGACAGTTCCGACCG	GAATATTTGAATAAAACAAAAGAATGAAT
rs4344834	ACGTTGGATGTGTGATGACATTGGGAGGTG	ACGTTGGATGAGCATTGGTCCGTTTGTGAG	aGTTTGTGAGCAGAGAGTCTCATT
rs6025931	ACGTTGGATGAAGTCTACGAAGCACTGAGC	ACGTTGGATGATGAATTTTGGAGGGACAC	gggTTTTGGAGGGACACTATCTT
rs7156868	ACGTTGGATGCTCATTTTTTCAATTTTAAAC	ACGTTGGATGAACCTCATCTGAGATGTTGGC	CCAAAATGATCTCACATTTCCAACAAGA
rs9521265	ACGTTGGATGACCTGGCCAAACATAACCTC	ACGTTGGATGGTCTGAATTCATTTTTCTC	gCATTTTTTCTCATGTATGTTGTTAAGA

## PLEX 4

SNP ID	Forward PCR primer	Reverse PCR primer	UEP primer
rs1006286	ACGTTGGATGATCACCAACTATGGTTTCCG	ACGTTGGATGAACACCGTTCACATCAGG	ATCAGGCCAGGATCTCC
rs16943226	ACGTTGGATGGTCACCAGAAAAATGCCTCC	ACGTTGGATGTGTTTATTGTGCTGGTTGGG	GGAGAAGGTGAACCC
rs2822757	ACGTTGGATGCCCAAATTTAGAGTCTTCTC	ACGTTGGATGGTGGTTCAGACAAAGGAACG	AACGGCCCAGTTGACTT
rs2883645	ACGTTGGATGAGATTAATGCCTGCTCCCTC	ACGTTGGATGTACAGGAGAGTTCTGGTTGG	AGTGCCCTTCTGGTA
rs1407467	ACGTTGGATGGTCGTAACCTACTCAACTTGTG	ACGTTGGATGGGAACAGAGCTACAACCTATTC	ACATGTTATCGATGGTTGC
rs17151653	ACGTTGGATGCAACTCCGAGGCTTGCTTTA	ACGTTGGATGATCAGCTGATGCCGTATTGC	CAGCACAGGTGGGGAGC
rs35806	ACGTTGGATGATGTGTTAAGCCCCACAGAG	ACGTTGGATGCCAGATACATCTGCCCAGTG	ATCTGCCCAGTGGCTCCC
rs6133535	ACGTTGGATGCTTAAAGGAAATGAGGAAGC	ACGTTGGATGCTGCCCCAAACTCTTTCATC	CTGTCTCTCCAAAGATAGC
rs1708759	ACGTTGGATGTGTGAATCCGCAACATTGGG	ACGTTGGATGAAATAAAAAATAGCTTCGTG	AAAATAGCTTCGTGTTTTGGTC
rs17237765	ACGTTGGATGAAGCATTGCTACCCATGTCC	ACGTTGGATGTGAGGCTGCAGCCATGAGAA	AGAAGGCCTGGCTGAGGAGCCCT
rs292482	ACGTTGGATGGTTGGAAGGTAGAATTTATC	ACGTTGGATGTCCTTACCCACACCTTCCTTG	TTTTTTTTTTCTGCATATCTGAA
rs4596920	ACGTTGGATGTACCTACTGTGTGAACCTACC	ACGTTGGATGTCCTTGTAGAATGTAGTCTC	TGTGTCAAATTGCTTTGTCC
rs10189158	ACGTTGGATGGCCATTTCAATAAGAACCA	ACGTTGGATGGAGCCTACCTTATTTCCAAG	ATTTCCAAGATCTTGAACCTGAAAT
rs10922106	ACGTTGGATGAGATTAGTTCATATTTATTG	ACGTTGGATGTCCTTCATTTATGAAGTTTAG	TCATTTATGAAGTTTAGTTTAGCA
rs12674544	ACGTTGGATGAAACATGATGACCCTAAAA	ACGTTGGATGCCCATGAATCATTAAATGGT	TATTTTGAATGAAATACAAAATAAACAA
rs12928389	ACGTTGGATGTCACCTCAATTGCGGACAG	ACGTTGGATGACGAGTCTAAATCTGGCACG	TAAATCTGGCACGCAGGTTAAGGAT
rs202916	ACGTTGGATGCTTCCTGCTATTTCAGTTTAG	ACGTTGGATGAGTGCCACCATAATCCACTC	AAGAGGAATATAATTCAAGAAACAGT

**Table A8.** Genomic positions (NCBI build36/hg18), size, number of probes, max log BF, and affected genes of 43 rare heterozygous germline deletions. Each CNV is present in only one patient.

Chro region	Position	Length (bp)	No. of probes	Max. Log BF	Involved genes
1p34.1	chr1:45592076-45700435	108360	18	57.1361	<i>TESK2</i>
1p31.1	chr1:71094328-71116548	22221	52	213.048	<i>PTGER3</i>
1p22.1	chr1:92580922-92800950	220029	48	207.295	<i>RPAP2, GFI1, EVI5</i>
1q44	chr1:246404335-246638006 <sup>#</sup>	233672	132	198.607	<i>OR2M2, OR2M3, OR2M4, OR2T33, OR2T12, OR2M7, OR14C36, OR2T4, OR2T6, OR2T1</i>
2q22.1 <sup>§</sup>	chr2:137473187-137569168	95982	33	122.085	<i>THSD7B</i>
2q22.1	chr2:137484018-137615223	131206	51	193.92	<i>THSD7B</i>
2q24.1	chr2:158813339-158905017	91679	51	196.779	<i>CCDC148</i>
3p26.3	chr3:260287-291417	31131	32	82.4472	<i>CHL1-UTR</i>
3p21.3	chr3:45839485-45852716	13232	8	20.4535	<i>LZTFL1</i>
3q26.31	chr3:176596105-176753724	157620	72	336.103	<i>NAALADL2</i>
4q21.22	chr4:83794836-83805695 <sup>#</sup>	10860	9	32.9419	<i>SCD5</i>
4q22	chr4:91075998-91166816	90819	20	92.0585	<i>MMRN1</i>
4q32.2	chr4:162981340-163113553	132214	38	97.0885	<i>FSTL5</i>
4q34.1	chr4:172930640-172979152	48513	12	21.5379	<i>GALNTL6</i>
5p15.32	chr5:5229672-5280314	50643	46	182.73	<i>ADAMTS16</i>
5q11.2	chr5:58578074-58734281	156208	72	326.102	<i>PDE4D</i>
5q22.2	chr5:112323863-112351637	27775	11	21.2786	<i>DCP2</i>
5q23.1	chr5:119920805-120247470	326666	107	609.173	<i>PRR16</i>
9p22.1	chr9:19101637-19117877 <sup>#</sup>	16241	12	48.4789	<i>ADFP</i>
9p21.3	chr9:20856974-20934558	77585	30	122.794	<i>FOCAD</i>
9p21.2	chr9:28465532-28493578 <sup>#</sup>	28047	16	61.4615	<i>LINGO2-UTR</i>
9q22.31	chr9:94670184-94740522	70339	24	60.0938	<i>ZNF484</i>
10q22.1	chr10:73839878-73855644	15767	10	26.3311	<i>CBARA1</i>
10q22.3	chr10:76856600-77074588	217989	50	117.305	<i>C10orf11</i>
11p15.4	chr11:8116419-8185496	69078	52	217.569	<i>RIC3</i>
11p11.2	chr11:44042462-44061557	19096	21	34.4747	<i>ACCS</i>
11q22.3	chr11:102695264-102848464 <sup>#</sup>	153201	64	209.884	<i>DYNC2H1</i>
11p11.2	chr11:44042462-44061557	19096	21	34.4747	<i>ACCS</i>
11q22.3	chr11:102695264-102848464 <sup>#</sup>	153201	64	209.884	<i>DYNC2H1</i>
12p12	chr12:20897382-21303121	405740	353	1606.22	<i>SLCO1B3, SLCO1B1</i>
12p12.2 <sup>§</sup>	chr12:20908555-20927800	19246	23	90.009	<i>SLCO1B3</i>
12q13.2	chr12:53445641-54110181 <sup>#</sup>	664541	256	1432.03	<i>MUCL1, KIAA0748, NEUROD4, OR9K1P, OR9K2, OR10A7, OR6C74, OR6C6, OR6C1, OR6C3, OR6C75, OR6C65, OR6C76</i>
13q31.1	chr13:84714710-85327835	613126	219	882.91	<i>SLITRK6</i>
14q31	chr14:87469000-87489111 <sup>#</sup>	20112	22	97.8503	<i>GALC</i>
15q15.1	chr15:40160584-40195058	34475	28	92.5148	<i>PLA2G4D</i>
16p13.3	chr16:3072198-3088992	16795	10	21.4463	<i>ZSCAN10</i>
17p13.2	chr17:5423583-5690836	267254	190	543.478	<i>NLRP1</i>

Chro region	Position	Length (bp)	No. of probes	Max. Log BF	Involved genes
17q12	chr17:29847372-30198038	350667	192	378.297	<i>C17orf102</i> , <i>TMEM132E</i>
18p11.21	chr18:14712857-14748540 <sup>#</sup>	35684	29	38.9638	<i>ANKRD30B</i>
19p13.12	chr19:15931793-15964505	32713	12	23.4107	<i>HSH2D-UTR</i>
19q13.43	chr19:62581812-62597371	15560	10	39.1491	<i>ZNF547-UTR</i> , <i>ZNF548</i>
19q13.43	chr19:62989408-63011738	22331	9	23.9233	<i>ZNF586</i> , <i>ZNF552</i>
Xq12	chrX:65699679-65934932 <sup>#</sup>	235254	23	150.713	<i>EDA2R</i>
Xq24	chrX:117506242-117629361 <sup>#</sup>	123120	30	181.128	<i>DOCK11</i>
Xq25	chrX:126827482-127023260	195779	29	58.5965	<i>ACTRT1</i>

<sup>#</sup> was found in 1 control; <sup>\$</sup> DNA from relatives of the index patient was available for co-segregation analysis. The CNV does not co-segregate with the phenotype.

**Table A9.** Genomic positions (NCBI build36/Hg18), size, number of probes, max log BF, and affected genes of 82 rare germline duplications

Chro region	Position	Length (bp)	No. of probes	Max. Log BF	Involved genes
1p32.3	chr1:54756695-54842677	85983	51	51.25	<i>ACOT11</i>
1p21.1	chr1:103166045-103200900	34856	27	54.4936	<i>COL11A1</i>
1q44	chr1:243187849-243498562	310714	140	260.502	<i>EFCAB2, KIF26B</i>
1q44	chr1:244601075-244774481	173407	108	189.364	<i>SMYD3, TFB2M</i>
2p16.1	chr2:56247422-56277102	29681	20	52.1844	<i>CCDC85A</i>
2q11.2	chr2:99205971-99283049	77079	29	68.9621	<i>LYG2, LYG1</i>
2q21.1	chr2:130539728-130864578	324851	108	147.542	<i>POTEF, CCDC74B, SMPD4, FAM128B, TUBA3E, CCDC115, IMP4, PTPN18</i>
2q35	chr2:220068927-220112970	44044	25	42.6615	<i>GMPPA, ACCN4, CHPF</i>
2q35	chr2:220125968-220188031	62064	38	59.2593	<i>OBSL1, INHA, STK11IP</i>
2q37.3	chr2:242322777-242393086	70310	40	31.1488	<i>D2HGDH, GAL3ST2</i>
3p26.3	chr3:1285741-1575422	289682	236	382.449	<i>CNTN6</i>
3p21.3	chr3:40985257-41512416	527160	187	305.322	<i>CTNBN1, ULK4</i>
3q23	chr3:143289345-143568000	278656	75	85.1647	<i>TFDP2, GK5, XRN1</i>
3q29 <sup>s</sup>	chr3:196055397-196587456	532060	230	393.975	<i>C3orf21, ACAP2</i>
3q29 <sup>s</sup>	chr3:197873977-198161806	287830	115	289.098	<i>C3orf34, PIGX, PAK2, SENP5, NCBP2, PIGZ</i>
4q31.21	chr4:144333863-144496222	162360	37	118.866	<i>USP38, GAB1</i>
5q23.1	chr5:118487627-118649571	161945	57	104.324	<i>DMXL1</i>
5q35.3 <sup>s</sup>	chr5:180379341-180448334	68994	38	34.7147	<i>BTNL9</i>
6q14.2	chr6:84547600-84691015	143416	39	118.216	<i>RIPPLY2, CYB5R4</i>
6q14.3	chr6:84847263-85138189	290927	89	210.706	<i>MRAP2, KIAA1009</i>
6q16	chr6:96654375-96715587	61213	27	52.6145	<i>FUT9</i>
6q24.2	chr6:144506912-144807770	300859	98	212.135	<i>STX11, UTRN</i>
7p22.3	chr7:1353711-1531935	178225	89	195.713	<i>MICALL2, INTS1</i>
7p21	chr7:21649161-21908645	259485	207	509.627	<i>DNAH11, CDCA7L</i>
7p11.2	chr7:55095278-55433865	338588	180	440.403	<i>EGFR, LANCL2</i>
7q22.1	chr7:100207480-100244080	36601	31	51.5787	<i>ZAN, EPHB4</i>
8p21.1	chr8:175887-277897	102011	81	67.1141	<i>ZNF596</i>
8p21.1	chr8:27522232-27652542	130311	60	108.328	<i>CLU, SCARA3, CCDC25</i>
8q11.2	chr8:52376255-52818072	441818	150	388.467	<i>PXDNL</i>
9p24.2	chr9:3739196-3835784	96589	59	136.453	<i>GLIS3</i>
9p24.2	chr9:3931973-3995592	63620	47	83.717	<i>GLIS3</i>
9p24.2	chr9:4011386-4196769	185384	168	396.494	<i>GLIS3</i>
9p24.2	chr9:4201589-4366397	164809	142	253.682	<i>GLIS3</i>
9p24	chr9:4400854-4489426	88573	58	198.9	<i>SLC1A1</i>
9p24	chr9:4495544-4537288	41745	28	60.3107	<i>SLC1A1</i>
9p22.1	chr9:19316053-19376565	60513	37	82.1049	<i>DENND4C, RPS6</i>
9q34.11	chr9:131514006-131766096	252091	154	360.921	<i>PRRX2, PTGES, TOR1B, TOR1A, C9orf78, USP20, FBNP1</i>
9q34.3	chr9:139403162-139519804	116643	52	144.059	<i>EXD3, NOXA1, ENTPD8, NELF, PNPLA7</i>



Chro region	Position	Length (bp)	No. of probes	Max. Log BF	Involved genes
9q34	chr9:139973769-140061878	88110	29	31.916	<i>CACNA1B</i>
9q34	chr9:140089674-140151812	62139	33	59.4435	<i>CACNA1B</i>
10p12.1 <sup>s</sup>	chr10:27654698-27832689	177992	116	165.246	<i>PTCHD3</i>
10q11.2	chr10:53735827-53798419	62593	23	60.2727	<i>DKK1</i>
11p15.4	chr11:5359130-5388870	29741	42	99.1799	<i>HBG2, HBE1, OR51M1, OR51J1</i>
11p15.1	chr11:18357549-18396314	38766	23	43.0579	<i>LDHA, LDHC</i>
11q22.3	chr11:108254505-108319661	65157	25	41.9953	<i>DDX10</i>
12q13.13	chr12:95960-333509	237550	162	253.467	<i>IQSEC3, SLC6A12, SLC6A13, KDM5A</i>
12p13.33	chr12:1071807-1311398	239592	78	125.37	<i>ERC1</i>
12p12	chr12:18459223-18485035	25813	11	35.4792	<i>PIK3C2G</i>
12q13.13	chr12:51330003-51442331	112329	109	120.634	<i>KRT2, KRT1, KRT77</i>
13q11	chr13:24169434-24228338	58905	22	52.948	<i>ATP12A</i>
13q12.3	chr13:30626847-30772444	145598	47	53.1348	<i>HSPH1, B3GALT1</i>
14q11.2	chr14:22925866-22953123	27258	32	43.1433	<i>MYH6, MYH7</i>
14q12	chr14:23985223-24057519	72297	34	44.7893	<i>SDR39U1, CMA1</i>
14q12	chr14:31579294-31628653	49360	19	64.3384	<i>ARHGAP5, C14orf128</i>
14q24.1	chr14:68992305-69046721	54417	13	33.1456	<i>SLC39A9</i>
14q32.33	chr14:103716079-103873344	157266	69	71.4027	<i>KIF26A</i>
15q26.1	chr15:88664833-88752520	87688	40	60.2516	<i>ZNF774, IQGAP1</i>
15q26.3	chr15:98367631-98710098	342468	217	312.989	<i>ADAMTS17</i>
16p12.3 <sup>s</sup>	chr16:20465942-20630271	164330	48	71.4004	<i>ACSM2, ACSM1</i>
16p12.3	chr16:22980334-23025207	44874	19	52.6576	<i>USP31</i>
16q23.2	chr16:79757966-79888598	130633	127	256.56	<i>PKD1L2, BCMO1</i>
17p13.2	chr17:3708409-3747829	39421	35	69.2656	<i>CAMKK1, P2RX1</i>
17q25.3	chr17:77925586-77986847	61262	26	49.3379	<i>UTS2R, C17orf101, HEXDC</i>
18p11	chr18:723474-929338	205865	79	162.859	<i>YES1, ADCYAP1</i>
18p11.22	chr18:9066995-9126155	59161	23	52.5421	<i>NDUFV2</i>
18q12	chr18:30679496-30915415	235920	78	129.858	<i>DTNA, MAPRE2</i>
18q12	chr18:31152359-32009041	856683	279	470.034	<i>ZNF24, ZNF396, INO80C, GALNT1, C18orf21, RPRD1A, SLC39A6, ELP2</i>
18q12	chr18:32045011-32098386	53376	24	55.0338	<i>MOCOS</i>
18q23	chr18:75342653-75507409	164757	136	198.576	<i>NFATC1</i>
18q23	chr18:75515931-75559025	43095	20	64.4395	<i>CTDP1</i>
19q13.43	chr19:61150873-61219157	68285	67	92.4029	<i>NLRP8, NLRP5</i>
19q13.43 <sup>s</sup>	chr19:61413858-61898207	484350	298	316.716	<i>ZSCAN5A, ZNF582, ZNF583, ZNF667, ZNF471, ZFP28, ZNF470, ZNF71, ZNF835</i>
20p13	chr20:828964-979043	150080	166	183.098	<i>ANGPT4, RSPO4</i>
20p12.3	chr20:5352295-5718919	366625	149	272.438	<i>C20orf196</i>
21q22.3	chr21:46515232-46536901	21670	20	44.2269	<i>MCM3AP, YBEY</i>
21q22.3	chr21:46540022-46625020	84999	34	121.771	<i>YBEY, C21orf58, PCNT</i>
22q13.31	chr22:43955895-44001348	45454	40	43.1688	<i>NUP50, KIAA0930</i>

Chro region	Position	Length (bp)	No. of probes	Max. Log BF	Involved genes
Xp22.33	chrX:6016157-6169284	153128	37	87.8562	<i>NLGN4X</i>
Xp22	chrX:18689860-19105471	415612	72	85.1306	<i>PPEF1, PHKA2, GPR64</i>
Xp11.23	chrX:46443751-46604442	160692	24	33.1653	<i>SLC9A7, RP2</i>
Xq24	chrX:117664921-117848187	183267	41	35.3568	<i>DOCK11, IL13RA1, ZCCHC12</i>
Xq24 <sup>\$</sup>	chrX:118311796-118492016	180221	36	71.2488	<i>SLC25A43, SLC25A5</i>

<sup>#</sup> was found in 1 control; <sup>\$</sup> DNA from relatives of the index patient was available for co-segregation analysis. The CNV does not co-segregate with the phenotype.

**Table A10.** The list of candidate genes present in the Steiner tree (second network analysis) (see Figure 4.20)

Type of aberration	Genes	No. of gene
Deletion	<i>ACCS, ADFP, CBARA1, DCP2, DOCK11, DYNC2H1, EDA2R, EVI5, GFI1, KIAA0748, FOCAD, LZTFL1, MMRN1, NAALADL2, NLRP1, PDE4D, PRR16, RIC3, RPAP2, SCD5, TESK2, THSD7B, ZNF484, ZNF548, ZNF552, ZNF586</i>	26
Partial gene duplication	<i>ACAP2, ADAMTS17, ANGPT4, ARHGAP5, B3GALT1, C20orf196, YBEY, KIAA0930, CACNA1B, CCDC25, CCDC85A, CDCA7L, CHPF, CLU, COL11A1, CTDSP1, D2HGDH, DDX10, DENND4C, DMXL1, DNAH11, DOCK11, DTNA, EGFR, EPHB4, ERC1, EXD3, FNBP1, GAB1, GLIS3, HEXDC, HSPH1, IQGAP1, KDM5A, KIF26B, KRT2, LANCL2, LDHC, MAPRE2, MCM3AP, MOCOS, MYH6, NFATC1, NLGN4X, OBSL1, PCNT, PIK3C2G, PXDN1, RP2, SDR39U1, SLC1A1, SLC9A7, SMYD3, STK11IP, TFB2M, TFDP2, ULK4, USP31, USP38, XRN1, YES1</i>	61
Whole gene duplication	<i>ADCYAP1, C17orf101, C18orf21, C21orf58, C3orf21, C3orf34, C9orf78, CCDC115, CTNNB1, DKK1, ELP2, GALNT1, GK5, GMPPA, IL13RA1, IMP4, INHA, INO80C, INTS1, KIAA1009, LDHA, MICALL2, NCBP2, NDUFB2, NELF, NOXA1, PAK2, PHKA2, PIGX, PTCHD3, PTPN18, RPRD1A, RPS6, SCARA3, SENP5, SLC25A43, SLC39A6, STX11, TOR1A, TOR1B, USP20, ZCCHC12, ZFP28, ZNF24, ZNF396, ZNF471, ZNF667, ZNF71</i>	48

*CNTN6* and *PTGES* are present in the 1<sup>st</sup> network but not in the 2<sup>nd</sup> network analysis

**Table A11.** Twenty-five networks of 180 candidate genes (**bold**) and the genes involved in the networks as given by Ingenuity pathway analysis

Net work	Top functions	Genes in network	Score	No. of candidate genes
1	Cancer, Cellular Development, Organismal Injury and Abnormalities	<b>ADCYAP1</b> , Akt, <b>CLU</b> , <b>CTDP1</b> , <b>CTNNB1</b> , <b>DKK1</b> , <b>EGFR</b> , <b>EPHB4</b> , ERK, ERK1/2, <b>GAB1</b> , <b>GFI1</b> , <b>GPR64</b> , Histone h3, <b>INHA</b> , <b>IQGAP1</b> , <b>KDM5A (JARID1A)</b> , Lh, <b>MAPRE2</b> , Mek, <b>MICU1 (CBARA1)</b> , NFkB(complex), <b>PAK2</b> , <b>PDE4D</b> , <b>PHKA2</b> , PI3K(complex), <b>PTGES</b> , RNA polymeraseII, <b>RPS6</b> , <b>SMYD3</b> , <b>TFB2M</b> , <b>USP20</b> , Vegf, <b>YES1</b> , <b>ZNF24</b>	46	25
2	Cellular Development, Hematological System Development and Function, Hematopoiesis	<b>ADD2</b> , <b>ARHGAP5</b> , CCND1, CD44, <b>CMA1</b> , <b>COL11A1</b> , CSF1, DSE, <b>EVI5</b> , Focal adhesion kinase, GLI1, HDAC2, HNRNPA2B1, <b>HSPH1</b> , IDS, MAPK1, MEST, MNS1, MTMR1, <b>NFATC1</b> , NTS, PDLIM3, PGM2L1, PHACTR2, <b>PKD1L2</b> , PTBP3, <b>PTGER3</b> , RGS10, SBDS, <b>SCD5</b> , <b>SLC1A1</b> , TLE1, <b>UTRN</b> , VIL1, <b>ZFP28</b>	17	12
3	Cell Cycle, Cancer, Cell Death and Survival	<b>ABLIM1</b> , <b>ADAMTS17</b> , AMOTL2, ARHGEF2, ARPC4, <b>CAMKK1</b> , <b>CDCA7L</b> , CPSF3, CSTF2, DHCR7, <b>DOCK11*</b> , <b>DTNA</b> , E4F1, <b>ERC1</b> , F Actin, FDFT1, FSH, HMGCS1, <b>LDHA</b> , LSS, MAMLD1, MGEA5, MLH1, <b>MOCOS</b> , <b>NCBP2</b> , PFKM, RASAL2, RB1, RGS12, SH3BP4, <b>TFDP2</b> , TP53, TPD52L1, <b>XRN1</b> , YWHAG	15	11
4	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry	AKAP12, ATPBD4, <b>DCP2</b> , <b>DIDO1</b> , EPG5, EXTL2, FABP5, GRB2, GTF2H5, HNRNPM, IGF2BP3, IL8, <b>IL13RA1</b> , KAT5, LBR, <b>MICALL2</b> , MYO1E, <b>NLRP1</b> , NOP58, NOX3, <b>NOXA1</b> , NOXO1, NUPR1, <b>PCNT</b> , PF4, <b>PLIN2 (ADFP)</b> , PPARG, RUVBL2, <b>SENP5</b> , SRF, <b>STX11</b> , STXBP2, TUBGCP2, <b>USP31</b> , XRN2	13	10
5	Cardiac Hypertrophy, Cardiovascular Disease, Cellular Compromise	<b>ALOX15</b> , <b>PRRX2</b>	2	1
6	Cancer, Cell Cycle, Cell Death and Survival	<b>CDH1</b> , <b>SLC39A6</b>	2	1
7	Developmental Disorder, Hereditary Disorder, Cellular Development	<b>INO80C</b> , MLL	2	1
8	Cancer, Cell Death and Survival, Neurological Disease	<b>TGM2</b> , <b>TOR1B</b>	2	1
9	Cell Cycle, Cellular Function and Maintenance, Cell-To-Cell Signaling and Interaction	<b>CHRNA7</b> , <b>RIC3</b>	2	1
10	Cellular Development, Cellular Function and Maintenance, Developmental Disorder	<b>GLIS3</b> , <b>SUFU</b>	2	1
11	Cell Cycle, Embryonic Development, Cellular Development	<b>CCDC85A</b> , NANOG	2	1
12	Cancer, Cellular Movement, Connective Tissue Development and Function	<b>KIF26B</b> , NQO1	2	1
13	Molecular Transport, RNA Trafficking, Cell Death and Survival	<b>RAE1</b> , <b>ZNF552</b>	2	1
14	Organ Morphology, Cancer, Endocrine System Disorders	<b>CLDN7</b> , <b>XXYLT1 (C3orf21)</b>	2	1
15	Cancer, Carbohydrate Metabolism, Cardiovascular System Development and Function	<b>KIT</b> , <b>PTPN18</b>	2	1

Net work	Top functions	Genes in network	Score	No. of candidate genes
16	Tissue Morphology, Cancer, Cell Morphology	<i>Mir-10, STK11IP</i>	2	1
17	Cancer, Cell Morphology, Cellular Response to Therapeutics	<i>MCM3AP, TP73</i>	2	1
18	Gene Expression, RNA Post-Transcriptional Modification, Cell Death and Survival	<i>SCARA3, SYVN1</i>	2	1
19	Cellular Assembly and Organization, RNA Post-Transcriptional Modification, Hereditary Disorder	<i>NDUFV2, RBM5</i>	2	1
20	Hereditary Disorder, Ophthalmic Disease, Neurological Disease	<i>KHSRP, RP2</i>	2	1
21	Cancer, Cell Cycle, Cellular Development	<i>CEBPA, MMRN1</i>	2	1
22	Cellular Movement, Skeletal and Muscular System Development and Function, Hair and Skin Development and Function	<i>BARX2, DMXL1</i>	2	1
23	RNA Post-Transcriptional Modification, Embryonic Development, Organismal Development	<i>INTS1, POLR2C</i>	2	1
24	Connective Tissue Disorders, Dermatological Diseases and Conditions, Hematological System Development and Function	<i>DENND4C, TLR7</i>	2	1
25	Cardiac Arrhythmia, Cardiac Hypertrophy, Cardiovascular Disease	<i>MYH6, MYOCD</i>	2	1

**Table A12.** The functional annotations by Ingenuity pathway analysis (IPA) of candidate genes associated with colorectal cancer

Functional annotation	No.of candidate gene	Candidate genes	p-value
Cancer	91	ACAP2, ACCS, ACSM1, ADAMTS16, ADAMTS17, ANGPT4, ANKRD30B, ARHGAP5, ASIC4, BCMO1, C21orf58, CACNA1B, CAMKK1, CCDC85A, CDCA7L, CHPF, CLU, CNTN6, COL11A1, CTNNB1, DKK1, DMXL1, DNAH11, DOCK11, DTNA, DYNC2H1, EDA2R, EGFR, ELP2, EPHB4, EVI5, FSTL5, GAL3ST2, GALC, GFI1, GK5, HEXDC, HSPH1, IQGAP1, KDM5A, KIAA0930, KIAA1009, KIF26B, KRT2, LDHA, LDHC, MCM3AP, MMRN1, MOCOS, MYH6, NAALADL2, NDUFV2, NLGN4X, NLRP1, PAK2, PDE4D, PIGX, PIK3C2G, PKD1L2, PLIN2, PRR16, PRRX2, PTCHD3, PTGER3, PTGES, PXDNL, RPS6, RSP04, SCD5, SLC39A6, SLC9A7, SLC01B3, SLITRK6, STX11, TESPA1, TFB2M, TFDP2, THSD7B, TMEM132E, TOR1A, ULK4, USP31, UTRN, XRN1, YES1, ZNF24, ZNF470, ZNF484, ZNF667, ZNF835, ZSCAN5A	$1.67 \times 10^{-2}$
Colon cancer	14	ARHGAP5, CLU, CTNNB1, DNAH11, EGFR, EPHB4, FSTL5, GAL3ST2, KIF26B, PTGES, SLC01B3, YES1, ZNF24, ZNF470	$2.31 \times 10^{-2}$
Colon carcinoma	11	ARHGAP5, CLU, CTNNB1, DNAH11, EGFR, EPHB4, FSTL5, GAL3ST2, KIF26B, ZNF24, ZNF470	$5.73 \times 10^{-3}$
Colon adenocarcinoma	10	ARHGAP5, CLU, CTNNB1, DNAH11, EGFR, EPHB4, FSTL5, GAL3ST2, KIF26B, ZNF470	$2.25 \times 10^{-3}$
Colorectal adenoma	2	CLU, CTNNB1	$1.11 \times 10^{-2}$
Adenomatous polyposis coli	1	CTNNB1	$4.67 \times 10^{-2}$

**Table A13.** Twenty-seven genes causing specific phenotype, reported in OMIM, DAVID, and GENATLAS.

Genes	Chro	Type of CNV	Descriptions	Mutation type
<i>ADAMTS17</i>	15q26.3	partial gene duplication	Weill-Marchesani-like syndrome (AR)	FS-DEL, FS-INS, SS
<i>B3GALT1</i>	13q12.3	partial gene duplication	Peters-plus syndrome (PpS) (AR)	FS-DEL, CSS
<i>BCMO1</i>	16q23.2	whole gene duplication	Hypercarotenemia and vitamin A deficiency (AD) (DAVID)	MS
<i>COL11A1</i>	1p21.1	partial gene duplication	Stickler syndrome; Marshall syndrome	SS
<i>CTDP1</i>	18q23	partial gene duplication	Congenital cataract, Facial dysmorphism neuropathy syndrome (AR)	MS
<i>D2HGDH</i>	2q37.3	partial gene duplication	D-2-hydroxyglutaric aciduria (AR)	MS, SS
<i>DNAH11</i>	7p21	partial gene duplication	Ciliary dyskinesia, primary, 7, with or without situs inversus	FS-INS, SS
<i>DTNA</i>	18q12	partial gene duplication	Noncompaction of left ventricular myocardium with congenital heart defects, Muscle dystrophies (AD)	MS
<i>DYNC2H1</i>	11q22.3	deletion	Asphyxiating thoracic dystrophy type 3 (ATD3)(AR)	MS
<i>GALC</i>	14q31	deletion	Globoid cell leukodystrophy (Krabbe disease) (AR)	MS
<i>GFI1</i>	1p22	deletion	Autosomal dominant severe congenital neutropenia	unknown
<i>KRT2</i>	12q13.13	partial gene duplication	Epidermal ichthyosis bullosa of Siemens	MS
<i>LDHA</i>	11p15.1	whole gene duplication	Glycogen storage disease XI with loss-of-function mutations. Exertional myoglobinuria due to deficiency of LDH-A (DAVID)	MS, FS-DEL
<i>MOCOS</i>	18q12	partial gene duplication	Xanthinuria type 2	MS
<i>MYH6</i>	14q11.2	partial gene duplication	Atrial septal defect 3, hypertrophic cardiomyopathy, familial, 6	MS
<i>NDUFV2</i>	18p11.22	whole gene duplication	Mitochondrial complex I deficiency; Hypertrophic cardiomyopathy with encephalopathy	MS, FS-DEL
<i>NELF</i>	9q34.3	whole gene duplication	Hypogonadotropic hypogonadism 9 with or without anosmia (AD)	MS
<i>OBSL1</i>	2q35	partial gene duplication	Gloomy face syndrome 2; 3M syndrome type 2 (AR)	MS, FS-DEL
<i>PCNT</i>	21q22.3	partial gene duplication	Microcephalic osteodysplastic primordial dwarfism, type II, truncating mutation; early onset (AR)	FS-DEL, NS
<i>PDE4D</i>	5q11.2	deletion	Acrodysostosis 2 (AD)	MS
<i>PHKA2</i>	Xp22	whole gene duplication	Glycogen storage disease type 9A (GSD9A) also known as X-linked liver glycogenosis (XLG)	NS, MS, CSS
<i>PKD1L2</i>	16q23.2	partial gene duplication	Conduct disorder and ADHD	unknown
<i>RP2</i>	Xp11.23	partial gene duplication	Retinitis pigmentosa-2, Xp11.3 deletion syndrome,	unknown
<i>RSPO4</i>	20p13	whole gene duplication	Congenital anonychia (AR)	MS
<i>SLCO1B3</i>	12p12	deletion	Hyperbilirubinemia (AR)	NS, DEL
<i>STX11</i>	6q24.2	whole gene duplication	Familial hemophagocytic lymphohistiocytosis (AR)	MS, FS-DEL, NS, DEL
<i>TOR1A</i>	9q34.11	whole gene duplication	Dystonia 1, idiopathic torsion (AD)	IF-DEL, MS

AD, autosomal dominant disease; AR, autosomal recessive disease; CSS, canonical splice site mutation; DEL, deletion; FS-DEL, frameshift deletion; FS-INS, frameshift insertion; IF-DEL, in-frame deletion; MS, missense mutation; NS, nonsense mutation, SS, splice site mutation



**Table A14.** The list of 54 from 180 candidate genes shown in STRING associated with known somatic mutated cancer genes reported in COSMIC database

<b>Genes</b>	<b>Chro</b>	<b>Type</b>	<b>Known somatically mutated cancer gene (COSMIC)</b>
<i>ADFP</i>	9p22.1	deletion	<i>ARNT</i>
<i>C10orf11*</i>	10q22.3	deletion	<i>CTNNB1</i>
<i>EDA2R</i>	Xq12	deletion	<i>TP53</i>
<i>FSTL5</i>	4q32.2	deletion	<i>BCR</i>
<i>GFI1</i>	1p22	deletion	<i>MYC</i>
<i>FOCAD</i>	9p21.3	deletion	<i>ASPSCR1</i>
<i>LZTFL1</i>	3p21.3	deletion	<i>CDH1</i>
<i>PDE4D</i>	5q11.2	deletion	<i>PDE4DIP</i>
<i>PRR16</i>	5q23.1	deletion	<i>TAL1</i>
<i>SLITRK6</i>	13q31.1	deletion	<i>NTRK1</i>
<i>TESK2</i>	1p34.1	deletion	<i>MUTYH</i>
<i>ZSCAN10</i>	16p13.3	deletion	<i>SOX2, POU5F1</i>
<i>ARHGAP5</i>	14q12	partial gene duplication	<i>ras homolog (RHO) family</i>
<i>CCDC25</i>	8p21.1	partial gene duplication	<i>SOS1-KRAS</i>
<i>CDCA7L</i>	7p15.3	partial gene duplication	<i>PHOX2B</i>
<i>CNTN6</i>	3p26.3	partial gene duplication	<i>NOTCH1, NOTCH2</i>
<i>CTDP1*</i>	18q23	partial gene duplication	<i>TCEA1</i>
<i>CYB5R4</i>	6q14.2	partial gene duplication	<i>DDIT3</i>
<i>D2HGDH</i>	2q37.3	partial gene duplication	<i>IDH1, IDH2</i>
<i>DDX10</i>	11q22.3	partial gene duplication	<i>HOXD13, WHSC1L1, RAP1GDS1, NUP98, HOXA9</i>
<i>EGFR</i>	7p11.2	partial gene duplication	<i>ERBB2, CBL</i>
<i>EPHB4</i>	7q22.1	partial gene duplication	<i>HOXA9, PIK3R1, PIK3CA</i>
<i>FNBP1*</i>	9q34	partial gene duplication	<i>DNM2</i>
<i>GAB1</i>	4q31.21	partial gene duplication	<i>EGFR, MET, PTPN11, PIK3R1</i>
<i>GLIS3</i>	9p24.2	partial gene duplication	<i>WWTR1</i>
<i>IQGAP1</i>	15q26.1	partial gene duplication	<i>CTNNB1, CDH1, KDR</i>
<i>KDM5A</i>	12p13.33	partial gene duplication	<i>RB1, SUZ12, ERCC2</i>
<i>LANCL2</i>	7q31	partial gene duplication	<i>EGFR, PPARG</i>
<i>MAPRE2</i>	18q12.1	partial gene duplication	<i>APC, RBM15</i>
<i>NFATC1</i>	18q23	partial gene duplication	<i>JUN, IL2</i>
<i>PCNT*</i>	21q22.3	partial gene duplication	<i>AKAP9</i>
<i>PIK3C2G</i>	12p12	partial gene duplication	<i>PTEN</i>
<i>PRRX2</i>	9q34.11	partial gene duplication	<i>JAK1, NUP98</i>
<i>SMYD3</i>	1q44	partial gene duplication	<i>MET, SET</i>
<i>STK11IP</i>	2q35	partial gene duplication	<i>STK11</i>
<i>TFDP2</i>	3q23	partial gene duplication	<i>RB1</i>
<i>YES1</i>	18p11	partial gene duplication	<i>CBLB, IL7R, KDR, EGFR, PDGFRB</i>
<i>CTNNB1</i>	3p21.3	whole gene duplication	<i>GSK3B, TCF7L2, AXIN1, CDH2, CTNNA1, APC, CDH1</i>
<i>DKK1</i>	10q11.2	whole gene duplication	<i>TP53, CTNNB1</i>
<i>GALNT1</i>	18q12.1	whole gene duplication	<i>MUC1</i>

<b>Genes</b>	<b>Chro</b>	<b>Type</b>	<b>Known somatically mutated cancer gene (COSMIC)</b>
<i>IL13RA1</i>	Xq24	whole gene duplication	<i>JAK1, JAK2, JAK3</i>
<i>INO80C</i>	18q12.2	whole gene duplication	<i>MLL</i>
<i>LDHA*</i>	11p15.1	whole gene duplication	<i>CREB1, MYC, SETD2, EP300, ARNT</i>
<i>MICALL2</i>	7p22.3	whole gene duplication	<i>HRAS</i>
<i>NELF*</i>	9q34.3	whole gene duplication	<i>FGFR1</i>
<i>RIPPLY2</i>	6q14.2	whole gene duplication	<i>CTNNB1, NOTCH1, NOTCH2</i>
<i>RPRD1A</i>	18q12.2	whole gene duplication	<i>CCND1, CCNE1</i>
<i>SCARA3</i>	8p21	whole gene duplication	<i>CD79A</i>
<i>SEN5</i>	3q29	whole gene duplication	<i>NPM1</i>
<i>SLC39A6</i>	18q12.2	whole gene duplication	<i>GATA3</i>
<i>USP20</i>	9q34.11	whole gene duplication	<i>VHL, ERG</i>
<i>ZCCHC12</i>	Xq24	whole gene duplication	<i>CTNNB1, PML, CREB1, TCF7L2</i>
<i>ZNF24</i>	18q12	whole gene duplication	<i>FANCA</i>
<i>ZNF667</i>	19q13.43	whole gene duplication	<i>SDHC</i>

\* causes specific phenotype

**Table A15.** The list of candidate genes reported to be associated with colorectal cancer.

Gene	Chro	Type	Descriptions
<i>ADFP</i> ( <i>PLIN2</i> )	9p22.1	deletion	Adipophilin, product of this gene, is increased in CRC patients and it can be used as a plasma biomarker for the detection of early-stage CRC (Matsubara et al. 2011).
<i>EDA2R</i>	Xq12	deletion	<i>EDA2R</i> (XEDAR) is a member of the TNFR superfamily and a putative tumor suppressor gene (Tanikawa et al. 2009). It functions as a direct p53 target and downregulated in CRC tissues (Tanikawa et al. 2010).
<i>GFI1</i>	1p22	deletion	<i>GFI</i> binding site is affected by a 2-bp GA deletion rs67491583. This variant is identified as a functional variant in CRC-associated enhancer MYC-335 (Tuupanen et al. 2012).
<i>FOCAD</i> ( <i>KIAA1797</i> )	9p21.3	deletion	<i>FOCAD</i> is presented in candidate predisposing CNVs in familial and early-onset CRC patients (Venkatachalam et al. 2011). Not only reported in CRC but also reported as a potential additional driver of breast cancers (Natrajan et al. 2012) and a novel focal adhesion protein with tumor suppressor function in gliomas (Brockschmidt et al. 2012).
<i>PTGER3</i>	1p31.1	deletion	<i>PTGER3</i> is a member of the G-protein coupled receptor family. This protein is one of four receptors identified for prostaglandin E2 ( <i>PGE2</i> ). Changes in <i>PGE2</i> receptors and synthesis in cell populations of precancerous familial adenomatous polyposis (FAP) colonic mucosa is defined by in situ and in vitro techniques (Takafuji et al. 2001). It plays an important role in suppression of cell growth and its down-regulation enhances colon carcinogenesis at a later stage (Shoji et al. 2004).
<i>SLCO1B3</i>	12p12	deletion	<i>SLCO1B3</i> or <i>OATP1B3</i> is frequent overexpressed in colorectal adenocarcinomas, it's overexpression reduces transcription activity of p53 (Lee et al. 2008).
<i>CLU</i>	8p21-p12	partial gene duplication	Clusterin (CLU) is a pleiotropic protein, plays an oncogenic role in colorectal tumorigenesis and progression (Xie et al. 2005). sCLU (a form of CLU) is overexpressed in CRC (Rodriguez-Pineiro et al. 2006). It could be a molecular marker for colon cancer screening (Chen et al. 2003; Mazzarelli et al. 2009).
<i>COL11A1</i>	1p21.1	partial gene duplication	<i>COL11A1</i> is up-regulated in most sporadic colorectal carcinomas (Fischer et al. 2001; Kim et al. 2010). Suceveanu et al. (2009) found mutation in <i>COL11A1</i> (exon 54) in colorectal tumor samples.
<i>EGFR</i>	7p11.2	partial gene duplication	<i>EGFR</i> is commonly overexpressed in CRC (Moran et al. 2004), activates in colorectal tumor (Hayashi et al. 1994; Karameris et al. 1993), and is one of drug target for CRC treatment (Bertotti et al. 2011; Prenen et al. 2013). <i>EGFR</i> copy number increases progressively in human colorectal carcinogenesis (Flora et al. 2012).
<i>EPHB4</i>	7q22.1	partial gene duplication	Inactivation of a single allele of <i>EphB4</i> results in higher proliferation in both the normal epithelium and intestinal tumors (Dopeso et al. 2009). This gene functions as a TSG (Dopeso et al. 2009; Ronsch et al. 2011).
<i>GAB1</i>	4q31.21	partial gene duplication	Overexpression of <i>GAB1</i> stimulates tumor progression in colorectal cancer cells (Moran et al. 2004; Seiden-Long et al. 2008).
<i>IQGAP1</i>	15q26.1	partial gene duplication	<i>IQGAP1</i> is a multifunctional protein involved in actin cytoskeleton assembly and E-cadherin-mediated cell adhesion (Shimao et al. 2002) and shows an over-expression in colorectal carcinoma tissues (Nabeshima et al. 2002). It plays a critical role in colon cancer cell invasion and it's high expression predicts poor prognosis in patients with colorectal carcinoma (Hayashi et al. 2010).
<i>MAPRE2</i>	18q12.1	partial gene duplication	<i>MAPRE2</i> or <i>RP1</i> is homolog to adenomatous polyposis coli-binding EB1-like gene family. Members of the EB1-like gene family may not only be involved in the tumorigenesis of colorectal cancers but may also play a role in the proliferative control of normal cells (Renner et al. 1997).
<i>PCNT</i>	21q22.3	partial gene duplication	<i>PCNT</i> aberration is found by array CGH in American African CRC patients (Brim et al. 2012).
<i>SMYD3</i>	1q44	partial gene duplication	<i>SMYD3</i> is overexpressed in CRC and validated as a biomarker for colorectal cancer (Xi et al. 2008). It has been reported the over-expression in colorectal tumor without <i>KRAS</i> mutation (Watanabe et al. 2011). This gene is also related to Wnt signaling pathway and other cancers (Hamamoto et al. 2004).
<i>YES1</i>	18p11	partial gene duplication	Increasing of c-YES activity may promote cancer spread and metastasis rather than tumor growth in colorectal carcinogenesis (Barraclough et al. 2007).

Gene	Chro	Type	Descriptions
<i>ADCYAP1</i>	18p11	whole gene duplication	<i>ADCYAP1</i> or <i>PACAP</i> not only acts as a neurotransmitter/neuromodulator but also plays a protective role in inflammatory bowel disease (IBD) and IBD-associated colorectal cancer in mice. <i>PACAP</i> knock out mice developed colorectal tumors with an aggressive-appearing pathology (Nemetz et al. 2008). <i>PACAP</i> is expressed in human colonic adenocarcinoma cell lines (Lelievre et al. 1998) and might play a role in the regulation of colon cancer growth via the Fas-R/Fas-L apoptotic pathway (Le et al. 2002).
<i>BCMO1</i>	16q23.2	whole gene duplication	Down regulation of <i>BCMO1</i> increases <i>ALDH1A</i> expression which activates PPARA pathway. The PPARA pathway influences oxidative damage and alters the expression of tumor suppressors which may contribute to intestinal tumorigenesis (Leclerc et al. 2013).
<i>CTNNB1</i>	3p21.3	whole gene duplication	<i>CTNNB1</i> , or Beta-Catenin, is a known cancer gene involves in many types of cancers reported in Cancer Gene Census (The Cancer Genome Project: CGP). It is one of the most common initially altered in sporadic colorectal tumors (Pendas-Franco et al. 2008). It functions as a key downstream oncogene, an overexpression of the gene could lead to tumorigenesis (Christie et al. 2013; Priolli et al. 2013).
<i>DKK1</i>	10q11.2	whole gene duplication	all known <i>DKK</i> genes were frequently silenced in colorectal cancer (CRC) cells but not in normal colon mucosa and that loss of DKKs may facilitate tumorigenesis through beta-catenin/T-cell factor-independent mechanisms (Sato et al. 2007). <i>DKK1</i> up-regulation acts as TSG (Pendas-Franco et al. 2008).
<i>GAL3ST2</i>	2q37.3	whole gene duplication	<i>GAL3ST2</i> is down-regulated in non-mucinous adenocarcinoma (Seko et al. 2002).
<i>LDHA</i>	11p15.1	whole gene duplication	<i>LDHA</i> is up-regulated in colorectal cell lines (Rutzky and Siciliano 1982) and in metastatic colorectal cancer (mCRC) patients with high serum lactate dehydrogenase (LDH) (Azuma et al. 2007).
<i>NOXA1</i>	9q34.3	whole gene duplication	<i>NOXA1</i> functions in the production of reactive oxygen species (ROS) and highly expressed in colon (Gianni et al. 2010). It activates NOX1 which highly expressed in human colon carcinoma cell line (Gianni et al. 2008).
<i>PAK2</i>	3q29	whole gene duplication	<i>PAK2</i> upregulation indicates acquisition of motility of human colon cancer C85 cells (Dabrowska et al. 2011).
<i>ZNF24</i>	18q12	whole gene duplication	<i>ZNF24</i> is located in chromosomal region that has been deleted in colorectal neoplasia (Rousseau-Merck et al. 1991).

**Table A16.** List of remaining 98 candidate genes after prioritization, for the point mutation screening

No	Gene	Chr.	Type	Affected part of gene	No.of patient	No. of control	Functions and pathways/Literature	No. of tumor with somatic mutations/total no. of tumors (%) **	Intolerance score <sup>§</sup> (percentile)	Haplo-insufficiency score (%) <sup>§§</sup>
<b>Heterozygous deletions</b>										
1	<i>ACCS</i>	11p11.2	DEL	5'UTR + exons 1-13	1	0	catalytic activity	19/636 (3.0)	72	75
2	<i>ADAMTS16</i>	5p15.32	DEL	exons 4-11	1	0	upregulated in esophageal squamous cell carcinoma and breast cancer	37/635 <b>(5.8)</b>	85	62
3	<i>C10orf11</i>	10q22.3	DEL	5'UTR + exons 1-2	1	0	<b>Wnt/beta-catenin signalling pathway</b>	4/610 (0.7)	66	81
4	<i>CBARA1</i>	10q22.1	DEL	exon 10	1	0	key regulator of mitochondrial calcium uptake	7/610 (1.2)	<b>24</b>	46
5	<i>CCDC148</i>	2q24.1	DEL	exons 5-10	1	0	<b>pathways in cancer, focal adhesion</b> , related to GI disease	16/599 (2.7)	91	80
6	<i>DCP2</i>	5q22.2	DEL	5'UTR + exons 1-2	1	0	RNA degradation	10/610 (1.6)	<b>10</b>	92
7	<i>DOCK11</i> *	Xq24	DEL/D UP	whole gene/exon 41-53 + 3'UTR	<b>2</b>	1 DEL	<b>regulate cell growth and differentiation</b> , gene family ( <i>DOCK1-11</i> ) involved in cancer	29/610 (4.7)	47	26
8	<i>EDA2R</i>	Xq12	DEL	whole gene	1	1 DEL	involved in <b>cell differentiation</b> , apoptosis, tumor necrosis factor-mediated signaling pathway	8/611 (1.3)	79	99
9	<i>EVI5</i>	1p22.1	DEL	exon 18 + 3'UTR	1	0	<b>regulator of cell cycle progression</b>	23/634 (3.6)	66	35
10	<i>FOCAD (KIAA1797)</i>	9p21.3	DEL	exons 20-30	1	0	<b>potential TSG in glioma/related to early-onset CRC?</b>	38/610 <b>(6.2)</b>	39	89
11	<i>FSTL5</i>	4q32.2	DEL	exon 4	1	0	<b>potential TSG?/calcium-ion binding</b>	43/636 <b>(6.7)</b>	81	53
12	<i>GALNTL6</i>	4q34.1	DEL	5'UTR + exon 1	1	0	transferase activity/Mucin type O-Glycan biosynthesis	23/635 (3.6)	<b>11</b>	n/a
13	<i>GFI1</i>	1p22	DEL	whole gene	1	0	<b>involved in regulation as a transcription repressor of G1/S phase of mitotic cell cycle</b>	16/610 (2.6)	73	30
14	<i>KIAA0748</i>	12q13.2	DEL	whole gene	1	1 DEL	positive regulation of T cell receptor signaling pathway	16/599 (2.6)	95	45
15	<i>LZTFL1</i>	3p21.3	DEL	exons 3-10 + 3'UTR	1	0	candidate cancer gene, <b>located on a hotspot for TSG</b>	5/610 (0.8)	80	74
16	<i>MMRN1</i>	4q22	DEL	exons 6-8 + 3'UTR	1	0	hemostasis, platelet degranulation, platelet activation	25/610 (4.1)	74	66
17	<i>NAALADL2</i>	3q26.31	DEL	exons 5-8	1	0	unknown	12/599 (2)	94	49
18	<i>NLRP1</i>	17p13.2	DEL	5'UTR + exons 1-3	1	0	<b>induction of apoptosis/NOD-like receptor signaling</b>	33/635 <b>(5.2)</b>	99	92
19	<i>PDE4D</i>	5q11.2	DEL	5'UTR + exons 1	1	0	<b>novel TSG</b> in esophageal adenocarcinoma	11/611 (1.8)	<b>4</b>	50

No	Gene	Chr.	Type	Affected part of gene	No.of patient	No. of control	Functions and pathways/Literature	No. of tumor with somatic mutations/total no. of tumors (%) **	Intolerance score <sup>s</sup> (percentile)	Haplo-insufficiency score (%) <sup>ss</sup>
20	<b>PLA2G4D</b>	15q15.1	DEL	5'UTR + exons 1-11	1	0	play a critical role in inflammation in psoriatic lesions	14/634 (2.2)	96	75
21	<b>PRR16</b>	5q23.1	DEL	exon 3 + 3'UTR	1	0	unknown	9/599 (1.5)	81	55
22	<b>PTGER3</b>	1p31.1	DEL	exons 6-7 + 3'UTR	1	0	colorectal cancer metastasis signaling	17/611 (2.7)	63	35
23	<b>RPAP2</b>	1p22.1	DEL	exons 12-13 + 3'UTR	1	0	RNA polymerase II-associated protein	10/610 (1.6)	<b>18</b>	81
24	<b>SLITRK6</b>	13q31.1	DEL	whole gene	1	0	unknown	35/634 ( <b>5.5</b> )	<b>13</b>	28
25	<b>TESK2</b>	1p34.1	DEL	exons 2-3	1	0	<b>Serin/Threonin-Proteinkinase, 4 kb upstream of <i>MUTYH</i></b>	15/665 (2.2)	68	39
26	<b>THSD7B</b>	2q22.1	DEL	exons 3-5	<b>2</b>	0	unknown	51/599 ( <b>8.5</b> )	n/a	32
27	<b>TMEM132E</b>	17q12	DEL	whole gene	1	0	unknown	24/599 (4.0)	62	30
28	<b>ZNF484</b>	9q22.31	DEL	exons 1-2	1	0	unknown, may be involved in transcriptional regulation	13/610 (2.1)	54	77
29	<b>ZNF552</b>	19q13.43	DEL	exon 3 + 3'UTR	1	0	unknown, may be involved in transcriptional regulation	13/610 (2.1)	62	75
30	<b>ZNF586</b>	19q13.43	DEL	exons 2-3 + 3'UTR	1	0	unknown, may be involved in transcriptional regulation	6/610 (1.0)	60	98
31	<b>ZSCAN10</b>	16p13.3	DEL	whole gene	1	0	Embryonic stem (ES) cell-specific transcription factor	12/611 (1.9)	<b>17</b>	<b>22</b>
<b>Partial duplications</b>										
32	<b>ARHGAP5</b>	14q12	DUP	5'UTR + exon 1	1	0	<b>Focal adhesion</b>	48/610 ( <b>7.8</b> )	<b>20</b>	44
33	<b>C20orf196</b>	20p12.3	DUP	5'UTR + exons 1-2	1	0	unknown	3/610 (0.5)	87	90
34	<b>C22orf9</b>	22q13.31	DUP	exons 2-9 + 3'UTR	1	0	unknown	7/610 (1.1)	<b>9</b>	58
35	<b>CACNA1B</b>	9q34	DUP	exons 32-44 + 3'UTR	1	1 DUP	regulation of calcium ion transport	58/634 ( <b>9.1</b> )	n/a	25
36	<b>CCDC25</b>	8p21.1	DUP	exon 9 + 3'UTR	1	0	unknown	2/610 (0.3)	55	82
37	<b>CCDC85A</b>	2p16.1	DUP	5'UTR + exons 1-2	1	0	candidate imprinted gene	9/599(1.5)	57	73
38	<b>CDCA7L</b>	7p15.3	DUP	exon 10 + 3'UTR	1	0	inhibits monoamine oxidase A, prevent cell apoptosis	13/610 (2.1)	<b>7</b>	31
39	<b>CLU</b>	8p21-p12	DUP	5'UTR + exons 1-3	1	0	unknown but expression increases in CRC	18/611 (2.9)	65	68
40	<b>CNTN6</b>	3p26.3	DUP	exons 9-23 + 3'UTR	1	1 DEL	<b>cell adhesion</b> , Notch signaling pathway	47/636 ( <b>7.4</b> )	<b>10</b>	<b>23</b>
41	<b>CYB5R4</b>	6q14.2	DUP	5'UTR + exons 1-10	1	0	plays a critical role in insulin production	22/636 (3.4)	71	47

No	Gene	Chr.	Type	Affected part of gene	No.of patient	No. of control	Functions and pathways/Literature	No. of tumor with somatic mutations/total no. of tumors (%) **	Intolerance score <sup>s</sup> (percentile)	Haplo-insufficiency score (%) <sup>ss</sup>
42	<b>D2HGDH</b>	2q37.3	DUP	exons 2-10 + 3'UTR	1	1 DUP	cellular protein metabolic process, small molecule metabolic process	7/599 (1.1)	49	39
43	<b>DDX10</b>	11q22.3	DUP	exons 17-18 + 3'UTR	1	1 DUP	<i>NUP98-DDX10</i> fusion increases proliferation and self-renewal of primary human CD34+ cells	13/637 (2.0)	<b>22</b>	<b>20</b>
44	<b>DENND4C</b>	9p22.1	DUP	exons 9-26 + 3'UTR	1	1 DUP	unknown	23/610 (3.7)	<b>9</b>	27
45	<b>DMXL1</b>	5q23.1	DUP	exons 12-43 + 3'UTR	1	0	unknown	37/610 ( <b>6.0</b> )	<b>12</b>	<b>11</b>
46	<b>EGFR</b>	7p11.2	DUP	exons 2-28 + 3'UTR	1	1 DEL	<b>involved in cancer pathways, cell proliferation, cell adhesion</b>	182/3467 ( <b>5.2</b> )	<b>1</b>	<b>0</b>
47	<b>EPHB4</b>	7q22.1	DUP	exons 13-17 + 3'UTR	1	0	<b>act as TSG</b> , angiogenesis pathway	22/641 (3.4)	<b>3</b>	<b>20</b>
48	<b>EXD3</b>	9q34.3	DUP	exon 22 + 3'UTR	1	0	unknown	13/610 (2.1)	99	n/a
49	<b>GAB1</b>	4q31.21	DUP	5'UTR + exons 1-2	1	0	<b>ErbB signaling pathway, cellular growth response, transformation and apoptosis</b>	26/637 (4.1)	54	<b>13</b>
50	<b>GLIS3</b>	9p24.2	DUP	5'UTR + exons 1-2 + 2 exons + exons 10-11 + 3'UTR	1	0	transcription activator & repressor	25/610 (4.1)	52	<b>15</b>
51	<b>HEXDC</b>	17q25.3	DUP	5'UTR + exons 1-4	1	1 DUP	carbohydrate metabolic process	15/610 (2.4)	81	62
52	<b>HSPH1</b>	13q12.3	DUP	5'UTR + exons 1-4	1	0	positive regulation of NK T cell activation	11/610 (1.8)	<b>19</b>	<b>6</b>
53	<b>IQGAP1</b>	15q26.1	DUP	5'UTR + exons 1-2	1	0	<b>Adherens junction, interact with APC</b>	36/610 ( <b>5.9</b> )	<b>5</b>	<b>17</b>
54	<b>KDM5A (JARID1A)</b>	12p13.33	DUP	exons 8-28 + 3'UTR	1	0	transcriptional regulation, chromatin modification	33/611 ( <b>5.4</b> )	<b>14</b>	<b>14</b>
55	<b>KIF26B</b>	1q44	DUP	5'UTR + exons 1-2	1	0	<b>positive regulation of cell-cell adhesion</b>	27/599 (4.5)	<b>18</b>	26
56	<b>LANCL2</b>	7q31.1-q31.33	DUP	5'UTR + exons 1-3	1	0	involved in negative regulation of transcription	12/610 (1.9)	50	75
57	<b>LYG1</b>	2q11.2	DUP	exons 4-8 + 3'UTR	1	0	peptidoglycan catabolic process	7/599 (1.1)	50	93
58	<b>MAPRE2</b>	18q12.1	DUP	5'UTR + exons 1-2	1	0	play a role in <b>proliferative control</b> of normal cells	7/611 (1.1)	38	<b>6</b>
59	<b>MCM3AP</b>	21q22.3	DUP	exons 21-28 + 3'UTR	1	0	<b>inhibits DNA replication and cell cycle progression</b>	35/637 ( <b>5.5</b> )	<b>1</b>	60
60	<b>NFATC1</b>	18q23	DUP	exons 12-13 + 3'UTR	1	0	transcription factor, <b>non-canonical Wnt signaling pathway</b>	24/637 (3.7)	<b>2</b>	38
61	<b>PIK3C2G</b>	12p12	DUP	exons 16-17	1	0	play roles in signaling pathways involved in <b>cell proliferation, oncogenic transformation</b>	32/611 ( <b>5.2</b> )	83	61
62	<b>PNPLA7</b>	9q34.3	DUP	exons 10-35 + 3'UTR	1	0	regulation of adipocyte differentiation	24/610 (3.9)	65	79

No	Gene	Chr.	Type	Affected part of gene	No. of patient	No. of control	Functions and pathways/Literature	No. of tumor with somatic mutations/total no. of tumors (%) **	Intolerance score <sup>s</sup> (percentile)	Haplo-insufficiency score (%) <sup>ss</sup>
63	<i>PRRX2</i>	9q34.11	DUP	exons 2-4 + 3'UTR	1	0	play a role in fetal skin development	3/610 (0.5)	n/a	<b>22</b>
64	<i>PXDNL</i>	8q11.21-q11.22	DUP	5'UTR + exons 1-21	1	0	oxidation-reduction process, hydrogen peroxide catabolic process	21/610 (3.4)	99	99
65	<i>SDR39U1</i>	14q12	DUP	5'UTR + exon 1	1	0	catalytic activity	8/599 (1.3)	95	n/a
66	<i>STK11IP</i>	2q35	DUP	5'UTR + exons 1-23	1	0	<b>TGFB1 signaling pathway/reported related to Peutz-Jaghers syndromes</b>	10/669(1.5)	37	74
67	<i>TFDP2</i>	3q23	DUP	5'UTR + exons 1-3	1	1 DUP	<b>cell cycle progression, involved in G1 phase of mitotic cell cycle</b>	8/599 (1.3)	43	<b>2</b>
68	<i>ULK4</i>	3p21.3	DUP	exons 33-37 + 3'UTR	1	0	unknown	9/630 (1.4)	96	72
69	<i>USP31</i>	16p12.3	DUP	exons 4-16 + 3'UTR	1	0	may be involved in ubiquitin-dependent protein catabolism	26/611 (4.2)	49	32
70	<i>USP38</i>	4q31.1	DUP	exons 4-10 + 3'UTR	1	0	involved in ubiquitin-dependent protein catabolism	24/611 (3.9)	<b>5</b>	65
71	<i>UTRN</i>	6q24.2	DUP	5'UTR + exons 1-13	1	0	positive regulation of cell-matrix adhesion	54/610 ( <b>8.8</b> )	<b>3</b>	51
72	<i>XRN1</i>	3q23	DUP	exons 30-42 + 3'UTR	1	0	<b>involved in homologous recombination, meiosis, telomere maintenance, and microtubule assembly</b>	34/610 ( <b>5.5</b> )	31	<b>11</b>
73	<i>YBEY (C21orf57)</i>	21q22.3	DUP	5'UTR + exons 1-3	1	0	putative metalloprotease or ribonuclease activity	5/610 (0.8)	54	62
74	<i>YES1</i>	18p11	DUP	5'UTR + exons 1-10	1	0	<b>adherens junction</b> , involved in G2/M progression and cytokinesis	13/641 (2.0)	39	<b>3</b>
75	<i>ZNF596</i>	8p23.3	DUP	exons 2-6 + 3'UTR	1	0	transcription, DNA-dependent, regulation of transcription, DNA-dependent	10/610 (1.6)	72	63
<b>Whole gene duplications</b>										
76	<i>CAMKK1</i>	17p13.2	DUP	whole gene	1	0	involved in <b>regulating cell apoptosis</b> , promotes cell survival	7/641 (1.1)	<b>19</b>	<b>13</b>
77	<i>CCDC115</i>	2q21.1	DUP	whole gene	1	0	unknown	2/610 (0.3)	35	76
78	<i>CTNNB1</i>	3p21.3	DUP	whole gene	1	0	<b>Wnt signaling pathway, known cancer gene</b>	380/6059 ( <b>6.2</b> )	<b>18</b>	<b>0.1</b>
79	<i>DKK1</i>	10q11.2	DUP	whole gene	1	0	<b>act as TSG, play a role in Wnt signaling pathway</b>	11/611 (1.8)	77	43
80	<i>ELP2</i>	18q12.2	DUP	whole gene	1	0	may have a role in chromatin structure and transcription	14/637 (2.2)	93	<b>18</b>
81	<i>GAL3ST2</i>	2q37.3	DUP	whole gene	1	0	involved in tumor metastasis process	14/610 (2.3)	n/a	91



No	Gene	Chr.	Type	Affected part of gene	No. of patient	No. of control	Functions and pathways/Literature	No. of tumor with somatic mutations/total no. of tumors (%) **	Intolerance score <sup>s</sup> (percentile)	Haplo-insufficiency score (%) <sup>ss</sup>
82	<b>GALNT1</b>	18q12.1	DUP	whole gene	1	0	Mucin type O-Glycan biosynthesis, metabolism of proteins	5/610 (0.8)	26	<b>25</b>
83	<b>INO80C</b>	18q12.2	DUP	whole gene	1	1 DEL	involved in transcriptional regulation, DNA replication and probably DNA repair	3/610 (0.5)	64	n/a
84	<b>KIAA1009</b>	6q14.3	DUP	whole gene	1	1 DEL	functions in cell division regulating chromosome segregation and mitotic spindle assembly	29/610 (4.8)	92	29
85	<b>PTPN18</b>	2q21.1	DUP	whole gene	1	0	proetin stability pathway	9/611 (1.5)	37	89
86	<b>RIPPLY2 (C6orf59)</b>	6q14.2	DUP	whole gene	1	0	ossification, somitogenesis	4/610 (0.7)	51	n/a
87	<b>RPS6</b>	9p22.1	DUP	whole gene	1	1 DUP	plays an essential role in <b>cell growth and proliferation</b> , mTOR signaling pathway	6/610 (1.0)	30	n/a
88	<b>RSPO4</b>	20p13	DUP	whole gene	1	0	<b>Wnt receptor signaling pathway</b> , activator of the beta-catenin signaling cascade	5/610 (0.8)	72	51
89	<b>SCARA3</b>	8p21-p12	DUP	whole gene	1	0	response to oxidative stress, <b>potential TSG</b>	3/611 (0.5)	<b>6</b>	56
90	<b>ZFP28</b>	19q13.43	DUP	whole gene	1	0	plays a role in embryonic development, regulation of transcription	26/610 (4.3)	58	49
91	<b>ZNF470</b>	19q13.43	DUP	whole gene	1	0	unknown, may be involved in transcriptional regulation	16/610 (2.6)	79	52
92	<b>ZNF471</b>	19q13.43	DUP	whole gene	1	0	unknown, may be involved in transcriptional regulation	22/636 (3.5)	93	47
93	<b>ZNF582</b>	19q13.43	DUP	whole gene	1	1 DUP	unknown, may be involved in transcriptional regulation	14/634 (2.2)	27	61
94	<b>ZNF583</b>	19q13.43	DUP	whole gene	1	0	unknown, may be involved in transcriptional regulation	23/610 (3.8)	36	64
95	<b>ZNF667</b>	19q13.43	DUP	whole gene	1	0	unknown, may be involved in transcriptional regulation	19/611 (3.1)	62	57
96	<b>ZNF71</b>	19q13.43	DUP	whole gene	1	0	unknown, may be involved in transcriptional regulation	12/610 (2.0)	63	43
97	<b>ZNF835</b>	19q13.43	DUP	whole gene	1	0	unknown, may be involved in transcriptional regulation	27/599 (4.5)	79	n/a
<b>Other</b>										
98	<b>UBC</b>	12q24.3		identified by network analysis	0	0	<b>DNA repair, cell cycle regulation</b> , protein degradation,	12/610 (1.9)	<b>12</b>	n/a

\* present both as deletion and duplication; \*\* tumors of the large intestine, COSMIC database; \$ Petrovski et al., 2013; \$\$, Huang et al., 2010; DEL, deletion; DUP, duplication; n/a, not available

**Table A17.** Twenty-three missense mutations identified in a group of whole gene duplications

Gene	Chro	Position	Exon	Allele1	Allele2	cDNA	Protein	dbSNP
<i>CAMKK1</i>	17	3776716	12	A	C	c.1151T>G	p.F384C	
<i>CAMKK1</i>	17	3788648	2	G	A	c.334C>T	p.H112Y	rs140915354
<i>CTNNB1*</i>	3	41266829	5	T	G	c.500T>G	p.V167G	
<i>DKK1</i>	10	54074755	2	G	A	c.316G>A	p.A106T	rs141115379
<i>DKK1</i>	10	54076088	3	G	A	c.440G>A	p.R147Q	rs371367754
<i>ELP2</i>	18	33716276	3	C	G	c.224C>G	p.S75C	rs74438152
<i>ELP2</i>	18	33747114	18	A	G	c.1990A>G	p.S664G	rs151280482
<i>ELP2</i>	18	33750104	17	G	A	c.1945G>A	p.V649M	
<i>GAL3ST2</i>	2	242716387	1	G	A	c.17G>A	p.G6D	rs117755329
<i>GAL3ST2</i>	2	242743565	4	C	G	c.1181C>G	p.P394R	
<i>KIAA1009</i>	6	84862387	23	A	G	c.3506T>C	p.V1169A	
<i>KIAA1009</i>	6	84913772	7	G	A	c.614C>T	p.P205L	
<i>PTPN18*</i>	2	131116862	3	A	C	c.259A>C	p.I87L	
<i>ZFP28</i>	19	57065068	8	A	T	c.914A>T	p.H305L	rs149264851
<i>ZFP28</i>	19	57065202	8	G	A	c.1048G>A	p.A350T	
<i>ZFP28</i>	19	57066595	8	A	T	c.2441A>T	p.N814I	
<i>ZNF471</i>	19	57035935	5	T	G	c.499T>G	p.C167G	
<i>ZNF471</i>	19	57036139	5	G	A	c.703G>A	p.A235T	rs377656535
<i>ZNF471*</i>	19	57036217	5	C	T	c.781C>T	p.L261F	
<i>ZNF471*</i>	19	57036986	5	G	A	c.1550G>A	p.C517Y	rs143533715
<i>ZNF667</i>	19	56953108	7	T	C	c.1256A>G	p.E419G	
<i>ZNF667</i>	19	56954077	7	G	A	c.287C>T	p.P96L	
<i>ZNF835</i>	19	57174975	2	G	A	c.1592C>T	p.P531L	

\* missense mutations predicted to be disease causing by all 3 in-silico tools (PolyPhen-2, Mutation Taster, SIFT)

**Table A18.** Logistic regression association of 117 SNPs from the replication GWAS study

SNP	Chr	Position	Ref allele	No. of sample	OR	SE	L95	U95	p-value	Bonferroni correction
rs4236978	8	56381051	T	895	1.375	0.112	1.105	1.711	0.004	0.501
rs797517	13	50128852	C	900	1.299	0.099	1.070	1.578	0.008	0.978
rs7156868	14	96898501	G	895	0.784	0.107	0.635	0.967	0.023	2.693
rs10501538	11	82464338	A	899	0.460	0.345	0.234	0.905	0.024	2.857
rs10823418	10	71109197	T	897	0.798	0.106	0.649	0.981	0.032	3.785
rs7727544	5	131618433	C	900	0.818	0.096	0.677	0.988	0.037	4.293
rs1354876	18	71820641	G	900	1.412	0.169	1.014	1.968	0.041	4.824
rs4858100	3	24057780	T	901	0.825	0.096	0.683	0.997	0.046	5.424
rs234434	14	96890773	G	895	0.818	0.105	0.666	1.004	0.055	6.431
rs17835866	14	96903431	A	898	0.817	0.107	0.662	1.008	0.060	6.971
rs9984896	21	36857539	C	898	1.240	0.122	0.976	1.576	0.078	9.149
rs1351805	14	96877218	C	898	0.833	0.105	0.678	1.023	0.082	9.537
rs9805437	13	101634297	G	902	1.493	0.250	0.914	2.439	0.109	12.776
rs2089855	5	131601428	C	900	1.166	0.097	0.964	1.410	0.113	13.268
rs10153396	18	29687410	A	901	1.235	0.135	0.948	1.608	0.117	13.736
rs798379	2	16755670	G	902	1.461	0.247	0.902	2.369	0.124	14.473
rs1708759	4	57539472	T	901	0.624	0.306	0.342	1.138	0.124	14.485
rs4344834	18	28400665	T	899	0.868	0.093	0.724	1.040	0.125	14.672
rs11955347	5	131595823	A	902	1.160	0.097	0.959	1.404	0.127	14.801
rs2400940	14	100294675	G	900	1.200	0.121	0.948	1.520	0.130	15.187
rs544276	9	112680289	C	899	1.205	0.128	0.939	1.548	0.144	16.790
rs4789409	17	72547423	A	899	0.854	0.109	0.690	1.056	0.145	16.965
rs330295	18	28403845	G	902	0.874	0.092	0.730	1.048	0.146	17.035
rs10848666	12	2479535	A	901	1.220	0.138	0.930	1.599	0.151	17.655
rs2965228	19	58442297	C	902	0.868	0.101	0.712	1.059	0.162	18.989
rs6061772	20	59738235	A	893	0.878	0.098	0.725	1.064	0.185	21.598
rs13357903	5	144235338	C	902	0.868	0.107	0.704	1.071	0.187	21.867
rs2304644	15	38532886	G	902	0.680	0.294	0.382	1.211	0.191	22.289
rs11578307	1	24325350	A	902	1.142	0.102	0.936	1.394	0.192	22.429
rs16943226	12	112717990	A	900	0.815	0.161	0.594	1.118	0.204	23.845
rs667808	10	78835668	G	901	0.883	0.101	0.724	1.075	0.215	25.108
rs60455014	14	7000655	C	900	1.422	0.298	0.794	2.548	0.236	27.659
rs17151639	7	127425052	G	894	0.880	0.109	0.711	1.089	0.238	27.893
rs2828064	21	23618125	G	900	0.873	0.116	0.695	1.096	0.242	28.349
rs11730575	4	178888021	A	901	0.876	0.115	0.699	1.098	0.250	29.262
rs1329428	1	194969433	A	902	0.895	0.100	0.736	1.088	0.264	30.900
rs11158362	14	61362431	G	902	0.694	0.336	0.359	1.340	0.277	32.351
rs2275199	1	155176319	A	902	0.870	0.132	0.672	1.127	0.293	34.246
rs624350	11	77501205	C	896	0.897	0.104	0.731	1.099	0.294	34.351
rs12674544	8	65857939	T	901	1.266	0.226	0.812	1.973	0.298	34.808
rs11629255	14	71670705	G	901	0.902	0.101	0.739	1.100	0.307	35.861
rs8192120	5	6685320	A	900	0.905	0.101	0.742	1.103	0.321	37.604
rs16932506	12	6272601	C	894	0.870	0.142	0.659	1.149	0.328	38.376
rs2274170	6	147227544	C	891	1.098	0.096	0.910	1.324	0.330	38.563
rs202916	22	33011163	C	901	0.913	0.096	0.756	1.102	0.342	40.026

SNP	Chr	Position	Ref allele	No. of sample	OR	SE	L95	U95	p-value	Bonferroni correction
rs443685	1	232629683	C	902	1.103	0.104	0.900	1.352	0.345	40.365
rs330297	18	28411147	T	902	1.098	0.099	0.904	1.334	0.346	40.482
rs2112613	5	114839566	G	890	0.903	0.109	0.729	1.118	0.347	40.611
rs11709614	3	24056776	A	898	0.914	0.097	0.756	1.105	0.353	41.266
rs10770675	12	20649080	A	900	0.899	0.114	0.719	1.125	0.353	41.278
rs9521265	13	108710327	A	891	1.107	0.109	0.894	1.370	0.353	41.278
rs389557	20	17763664	A	901	0.914	0.097	0.755	1.106	0.354	41.371
rs17151653	7	127433088	G	901	0.905	0.109	0.731	1.120	0.359	42.026
rs10065787	5	131464385	T	902	1.090	0.095	0.905	1.312	0.365	42.728
rs927596	13	99334817	T	900	1.125	0.132	0.869	1.456	0.371	43.372
rs35806	10	78835836	G	902	0.916	0.100	0.754	1.114	0.380	44.402
rs12139641	1	155152391	G	899	0.893	0.132	0.690	1.155	0.388	45.396
rs7997539	13	103627451	G	857	1.097	0.108	0.888	1.356	0.391	45.794
rs912284	13	99335233	T	901	1.116	0.131	0.862	1.443	0.405	47.362
rs292488	5	168041837	A	902	0.910	0.120	0.720	1.151	0.432	50.579
rs1020430	2	133409885	T	901	0.888	0.156	0.654	1.205	0.445	52.007
rs822431	1	155168905	G	902	0.907	0.128	0.707	1.165	0.447	52.276
rs16931374	8	65749351	T	901	1.184	0.234	0.749	1.872	0.469	54.873
rs8098464	18	28516811	G	849	0.901	0.147	0.676	1.201	0.477	55.751
rs2073167	22	40121482	C	890	0.937	0.096	0.776	1.130	0.495	57.915
rs1244229	10	8047566	T	872	1.075	0.110	0.867	1.334	0.509	59.600
rs4860701	4	66266097	T	896	1.068	0.099	0.879	1.297	0.511	59.799
rs4968046	16	22985890	T	902	0.923	0.125	0.723	1.179	0.521	60.969
rs16957347	16	10496762	T	901	0.897	0.172	0.640	1.256	0.525	61.472
rs10189158	2	175746125	A	902	1.147	0.223	0.742	1.775	0.537	62.771
rs942876	13	23512951	A	901	1.062	0.097	0.877	1.285	0.540	63.122
rs292489	5	167949578	T	901	0.936	0.110	0.754	1.161	0.546	63.835
rs4596920	1	159189375	T	902	1.063	0.102	0.871	1.297	0.548	64.093
rs12935619	16	23035790	T	901	0.928	0.125	0.727	1.186	0.552	64.584
rs2883645	15	22514468	T	901	1.065	0.111	0.856	1.324	0.572	66.924
rs7521700	1	168069384	G	901	1.056	0.098	0.872	1.279	0.576	67.369
rs7569783	2	5780740	C	902	1.064	0.118	0.845	1.340	0.600	70.153
rs3111779	2	231501259	T	901	1.077	0.145	0.811	1.431	0.610	71.323
rs11012878	10	22481893	C	902	0.922	0.160	0.673	1.262	0.610	71.370
rs17346550	1	171582248	C	894	0.949	0.105	0.772	1.166	0.616	72.107
rs1512414	6	13384394	T	900	0.946	0.117	0.752	1.191	0.638	74.646
rs149476	5	167939278	A	864	0.954	0.109	0.770	1.181	0.665	77.817
rs6536530	4	161790521	A	900	1.042	0.097	0.862	1.259	0.671	78.519
rs6025931	20	55968066	C	901	0.962	0.094	0.800	1.156	0.677	79.151
rs292482	5	167989258	T	891	0.954	0.114	0.763	1.193	0.678	79.268
rs12928389	16	10533623	C	902	1.074	0.172	0.767	1.504	0.678	79.303
rs7191411	16	10529441	A	896	1.073	0.173	0.765	1.505	0.682	79.771
rs12597756	16	10381318	A	902	0.933	0.179	0.657	1.324	0.697	81.549
rs1887432	9	36721326	G	902	0.948	0.144	0.715	1.258	0.712	83.304
rs6089491	20	59730129	T	899	0.966	0.098	0.797	1.170	0.720	84.193
rs6133535	20	7986649	G	902	1.069	0.193	0.733	1.560	0.727	85.106
rs2822757	21	14831122	C	902	1.036	0.107	0.840	1.279	0.739	86.510

SNP	Chr	Position	Ref allele	No. of sample	OR	SE	L95	U95	p-value	Bonferroni correction
rs1864262	2	5777922	G	898	0.971	0.093	0.808	1.166	0.750	87.703
rs8192166	5	6695356	T	897	1.032	0.100	0.848	1.255	0.754	88.241
rs731326	23	13438391	C	898	0.974	0.086	0.822	1.153	0.755	88.370
rs158896	5	167915925	T	900	0.967	0.111	0.779	1.201	0.764	89.341
rs13419910	2	5757137	G	899	1.028	0.101	0.844	1.254	0.781	91.412
rs1006286	10	119509000	C	901	0.949	0.205	0.635	1.418	0.798	93.343
rs195656	16	69604985	A	893	0.964	0.143	0.728	1.277	0.799	93.448
rs1799932	22	40241471	T	899	0.977	0.095	0.811	1.177	0.806	94.349
rs6570786	6	147395785	T	900	0.977	0.098	0.806	1.185	0.814	95.180
rs11161353	15	21655245	T	901	1.033	0.146	0.777	1.374	0.822	96.127
rs35407548	1	171593970	C	895	0.980	0.106	0.797	1.206	0.851	99.544
rs2918006	2	231492535	G	901	1.026	0.137	0.784	1.343	0.851	99.567
rs7920199	10	134078268	G	901	1.022	0.130	0.792	1.317	0.869	101.720
rs6884552	5	6714966	T	902	1.016	0.100	0.834	1.236	0.877	102.574
rs2169123	15	21677404	A	902	1.021	0.142	0.774	1.348	0.882	103.136
rs6981465	8	65612808	T	901	1.021	0.150	0.760	1.370	0.892	104.341
rs12015040	23	13437789	C	902	0.989	0.086	0.836	1.171	0.897	104.972
rs6475895	9	26443356	G	897	0.980	0.166	0.708	1.356	0.902	105.569
rs6997421	8	65643604	A	901	1.026	0.216	0.673	1.566	0.905	105.827
rs16931326	8	65657887	A	901	0.983	0.225	0.633	1.527	0.939	109.851
rs2158641	7	48891030	G	891	0.992	0.102	0.812	1.213	0.941	110.085
rs4392152	18	73915669	C	901	0.993	0.097	0.820	1.202	0.942	110.226
rs1407467	9	7958906	G	901	1.004	0.096	0.832	1.213	0.964	112.788
rs11645638	16	77517824	T	901	0.996	0.094	0.828	1.198	0.966	112.999
rs17237765	6	147378663	O	902	NA	NA	NA	NA	NA	NA

NA, not application; OR, odd ratio; SE, standard error